

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOSOFÍA

DEPARTAMENTO DE FILOSOFÍA DEL DERECHO, MORAL Y

POLÍTICA II (ÉTICA Y SOCIOLOGÍA)



TESIS DOCTORAL

Moralidad y cooperación racional

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Blanca Rodríguez López

DIRIGIDA POR

Gilberto Gutiérrez López

Madrid, 2002

R. 23.233

333/91

UNIVERSIDAD COMPLUTENSE DE MADRID

TES

Facultad de Filosofía y Ciencias de la Educación

Departamento de Filosofía del Derecho, Moral y Política II

(Ética y Sociología)

MORALIDAD Y COOPERACION RACIONAL



Blanca Rodríguez López

Madrid, 1991

Colección Tesis Doctorales. N.º 333/91

© Blanca Rodríguez López

Edita e imprime la Editorial de la Universidad
Complutense de Madrid. Servicio de Reprografía.
Escuela de Estomatología. Ciudad Universitaria.
Madrid, 1991.
Ricoñ 3700
Depósito Legal: M-42090-1991



La Tesis Doctoral de Dña Blanca Rodríguez.....
..López.....
Titulada "Moralidad y cooperación racional".....
.....
Director Dr. D. Gilberto Gutiérrez López.....
fue leída en la Facultad de Filosofía y CC. Edc....
de la UNIVERSIDAD COMPLUTENSE DE MADRID, el día .8....
de marzo de 19⁹¹....., ante el tribunal
constituido por los siguientes Profesores:
PRESIDENTE Dr. D. Victoriano Martín Martín.....
VOCAL Dr. d. José Montoya Sáenz.....
VOCAL Dr. D. Juan Miguel Palacios García.....
VOCAL Dr. D. Damián Salcedo Megalés.....
SECRETARIO Dr. D. Jaime Sarabia Alvarez-Ude.....
.....
habiendo recibido la calificación de *APTO cum*.....
..... LAUDE POR UNANIMIDAD.....

Madrid, a 8 de mayo de 1991.
EL SECRETARIO DEL TRIBUNAL.

arc

BLANCA RODRIGUEZ LOPEZ

MORALIDAD
Y COOPERACION RACIONAL

TESIS DOCTORAL

DIRECTOR:

Catedrático Doctor Don GILBERTO GUTIERREZ LOPEZ

DEPARTAMENTO DE FILOSOFIA DEL DERECHO,
MORAL Y POLITICA II (ETICA Y SOCIOLOGIA)
FACULTAD DE FILOSOFIA Y CIENCIAS DE LA EDUCACION
UNIVERSIDAD COMPLUTENSE DE MADRID

1990

INDICE

INTRODUCCION.	1-iv
PARTE PRIMERA. LA TEORIA	1
CAPITULO I. RACIONALIDAD PRACTICA	2
CAPITULO II. CONDICIONES DE RACIONALIDAD	13
1. Condiciones formales de racionalidad.	14
2. Condiciones materiales de racionalidad.	21
CAPITULO III. SITUACIONES DE ELECCION	33
1. Situaciones de certidumbre.	34
2. Situaciones de riesgo e incertidumbre.	36
3. Satisfacción y maximización.	45
CAPITULO IV. RACIONALIDAD ESTRATEGICA	50
SEGUNDA PARTE. EL PROBLEMA DE LA COOPERACION RACIONAL	71
CAPITULO V. ACUERDOS RACIONALES	73
CAPITULO VI. ¿ES RACIONAL MANTENER LOS ACUERDOS?	88
I. Acuerdos satisfactorios.	91
II. Por qué aceptar las condiciones de negociación.	109
CAPITULO VII. LA ESTABILIDAD DE LOS ACUERDOS SATISFACTORIOS	122
PARTE TERCERA. COMO HACER LOS ACUERDOS IMPONIBLES	134
CAPITULO VIII. LA DERIVACION DE LA MORAL A PARTIR DEL INTERES PROPIO	145
CAPITULO IX. LA MORAL COMO PUNTO DE VISTA	169
CUARTA PARTE. POR QUE ADOPTAR EL PUNTO DE VISTA MORAL	182
CAPITULO X. RACIONALIDAD INDIRECTA	185
I. Racionalidad directa y Racionalidad indirecta.	185
II. Qué es lo que nos pide nuestra teoría.	207
CAPITULO XI. UNA TEORIA ALTERNATIVA.	215
I.	216
II. La teoría de los objetivos presentes.	220
III. Los argumentos contra T.	227
IV.	256
CAPITULO XII. LA MORAL COMO RACIONALIDAD INDIRECTA	264

APENDICES	282
APENDICE 1.	283
APENDICE 2	294
APENDICE 3	299
APENDICE 4.	304
APENDICE 5	307
APENDICE 6.	313
APENDICE 7	318
APENDICE 8	333
APENDICE 9	336
APENDICE 10	346
 BIBLIOGRAFIA	 367
 INDICE DE AUTORES	 370
 INDICE DE MATERIAS	 371

INTRODUCCION

Hay una concepción de la racionalidad práctica o, si se prefiere, una teoría sobre la racionalidad práctica, bien definida, según la cual la conducta racional es aquella que maximiza la utilidad del agente, definida sobre el conjunto de sus preferencias racionales. Esta teoría es buena y resulta defendible por diversos motivos.

Sin embargo, en determinadas situaciones la aplicación de esta teoría da lugar al surgimiento de dilemas. Estos dilemas consisten básicamente en lo siguiente: hay ocasiones en las que si cada uno hace lo mejor según la teoría, el resultado es peor para todos de lo que sería si ninguno obedeciera la teoría.

Hay dos modos de solucionar estos dilemas. Por un lado, puede modificarse la situación de tal modo que desaparezcan las condiciones que dan lugar al dilema, con lo cual este deje de plantearse. Esto soluciona el problema práctico y evita el problema teórico, puesto que entonces cada uno de los agentes obraría de acuerdo con la teoría y conseguiría lo mejor. Por otro lado, está lo que se ha llamado "solución moral". La aplicación de esta solución también solventaría el problema práctico, pero no evitaría el problema teórico, pues los agentes conseguirían lo mejor al precio de desobedecer la teoría. Los agentes conseguirían lo que según la teoría es racional intentar conseguir, pero lo harían comportándose de un modo que esta teoría califica de irracional. El dilema seguiría existiendo.

La existencia de estos dilemas plantea un problema indiscutible a la teoría. Parece que debemos revisar sus supuestos e intentar modificarlos para obtener una teoría más satisfactoria. Concretamente, debemos revisar la afirmación de nuestra teoría de que la conducta moral es irracional. Esta afirmación es problemática en sí misma, independientemente del papel que desempeña en el surgimiento de los dilemas. Si modificamos esta afirmación, entonces los agentes que se comportan moralmente conseguirán lo mejor mediante una conducta racional.

Sin embargo, admitir la racionalidad de la conducta moral no evita el surgimiento de los dilemas, a menos que podamos también demostrar que la conducta moral está racionalmente exigida. Si no es así, para los agentes que en una situación de dilema se comportan de un modo no moral seguirá planteándose un dilema. Ya no podrá decirse que el resultado que alcanzan es peor del que obtendrían si desobedecieran a la teoría, pues

comportarse moralmente ya no puede considerarse como una conducta contraria a la teoría. Pero sí podrá decirse que estos agentes, comportándose de un modo racional, obtienen unos resultados subóptimos.

Este es fundamentalmente el problema a cuyo estudio está dedicada la presente tesis doctoral. La teoría acerca de la racionalidad práctica de la que se parte, si bien es satisfactoria en el nivel individual, resulta auto-frustrante a nivel colectivo: en determinadas situaciones sucede que, si bien cada uno de los agentes individuales implicados en la elección hace lo mejor que puede hacer, entre todos hacen algo que no es lo mejor posible. Esto contrasta con lo que sucedería si los agentes se comportaran en esas situaciones de un modo moral. Entonces los agentes individuales actuarían de un modo irracional, pero el resultado sería para todos el mejor posible. En esto precisamente consiste la paradoja de la mencionada teoría sobre la racionalidad práctica: que en algunas situaciones, si todos obramos racionalmente, de acuerdo con la teoría, cumpliremos los objetivos propuestos por la teoría peor de lo que si obráramos irracionalmente. Y el propósito del presente trabajo es analizar cómo y por qué surge esta paradoja e intentar resolverla sin tener que renunciar a una teoría de la racionalidad práctica que por lo demás resulta satisfactoria.

La tesis está estructurada en cuatro partes, cada una de las cuales está compuesta por varios capítulos. La primera parte está destinada a analizar la teoría de la racionalidad práctica y su aplicación a distintos tipos de situaciones. Básicamente, la teoría de la que se parte es la utilizada en la teoría económica¹. Esta teoría sólo se ocupa de lo que podríamos llamar el lado formal de la elección racional. Es decir, en la teoría económica se toman las preferencias e intereses del agente como dados y sólo se trata de analizar cómo este debe elegir, dado que tiene las preferencias que tiene. Sin embargo, esto no resulta muy satisfactorio, sino que más bien parece que es necesario admitir que algunas preferencias pueden resultar por sí mismas irracionales. Esto nos lleva a una ampliación o, mejor, a una complementación de la teoría con unos supuestos sobre la racionalidad de las preferencias. Una parte de esta amplia-

¹ Concretamente, la teoría utilizada aquí es básicamente la teoría Bayesiana. Sin embargo, la parte sustancial de la tesis que trata sobre los aspectos de la racionalidad

ción se basa en los trabajos de algunos autores, en los que se critican las preferencias por el modo en que estas surgen y se mantienen. Otra parte constituye una modificación de la teoría habitual como una respuesta a determinadas críticas².

Esta primera parte está dividida en cuatro capítulos. En el capítulo I se analizan las ideas preteóricas acerca de la racionalidad práctica y se muestra cómo nuestra teoría las recoge. En el capítulo II se establecen las condiciones que deben cumplir tanto las preferencias aisladas como el conjunto de preferencias de un individuo para que sea posible una elección racional. En el capítulo III se analiza el funcionamiento de la teoría en situaciones de riesgo e incertidumbre. Por último, en el capítulo IV se estudia la aplicación de la teoría a situaciones de naturaleza estratégica, en las cuales la obtención de un determinado resultado depende de la acción de varios agentes racionales.

En la segunda parte, dividida en tres capítulos, se estudia el problema de la cooperación racional, es decir, de las situaciones en las que a los agentes racionales les resulta mutuamente beneficioso cooperar. En el capítulo V se analiza el problema del regateo y sus soluciones, esto es, de cómo seleccionar una estrategia conjunta entre todas las posibles. El capítulo VI trata de la distinción entre acuerdos insatisfactorios y acuerdos satisfactorios y en el capítulo VII se plantea el problema de la estabilidad de los acuerdos satisfactorios y en qué condiciones es esta posible.

La tercera parte se centra en el problema de cómo garantizar el cumplimiento de los acuerdos cooperativos. En el capítulo VIII se analiza la posibilidad de derivar la conducta moral a partir del interés propio de un modo directo, es decir, de eliminar la contraposición entre los dictados de la moralidad y los del interés. Tras desechar tal posibilidad, el capítulo IX se plantea la moral como un punto de vista que un agente racional puede o no adoptar, a saber, aquel punto de vista que un agente tendría si se encontrara en determinada situación hipotética.

La cuarta y última parte analiza las razones que puede tener un agente racional para comportarse de un modo moral. El capítulo X establece, en primer lugar, la distinción entre racionalidad directa y racional-

² Concretamente, he realizado estas modificaciones, o mejor, matizaciones, de la teoría como respuesta a las críticas de Parfit.

lidad indirecta, así como su relación con el problema de la debilidad de la voluntad y analizando, en segundo lugar, en qué situaciones entra en juego la racionalidad indirecta, estableciendo que, en determinadas circunstancias, una teoría acerca de la racionalidad práctica puede pedir a los agentes racionales que se comporten de un modo irracional. En el capítulo XI se estudian las características de nuestra teoría que la hacen ser en ocasiones auto-frustrante a nivel colectivo, al tiempo que se analiza la posibilidad de una teoría alternativa sobre la racionalidad práctica que no presente este problema. Una vez desestimada esta posibilidad, en el capítulo XII y último se establece la posibilidad de considerar la conducta moral como un método de racionalidad indirecta.

Además de la estructura en cuatro partes que acabamos de presentar, la presente tesis contiene varios apéndices. En ellos se tratan diversos temas que, si bien no son esenciales para el seguimiento de la argumentación principal, son de interés por distintas razones. Algunos de ellos, como el apéndice 6, son simplemente ejemplos extensos. Otros, como por ejemplo el apéndice 1, son estudios de temas colaterales que me han parecido especialmente interesantes. Y todos ellos, aun sin ser esenciales, pueden contribuir a la comprensión de la argumentación principal.

PARTE PRIMERA

LA TEORIA

CAPITULO I

RACIONALIDAD PRACTICA

Los conceptos de racionalidad práctica y acción racional están en estrecha relación con el concepto de elección. En efecto, sólo en aquellas situaciones en las que es posible elegir uno entre varios cursos de conducta alternativos puede plantearse la cuestión de qué conducta entre las posibles es la que la razón nos aconseja seguir. Por este motivo, los conceptos de acción racional y de elección racional pueden tratarse como equivalentes.

Sin embargo, por razones metodológicas, es preferible utilizar el término "elección racional" al menos en principio, puesto que en él se aprecian con mayor claridad los aspectos más característicos del concepto.

No sólo es cierto que existe la posibilidad de elección. Es también cierto, e igualmente indiscutible, que la elección es en muchos casos no sólo una posibilidad, sino también una necesidad. Es decir, la elección es algo ineludible, y lo es en gran cantidad de casos de índole e importancia muy diversa; tenemos que elegir entre ir al cine y quedarnos en casa, entre votar a uno u otro partido político en las elecciones o no votar en absoluto, entre hacer una tesis doctoral o buscar un trabajo bien remunerado, uno de entre varios pretendientes.

En cada uno de estos casos, seguir un determinado curso de acción implica desechar otros. Es por tanto indiscutible que para actuar es necesario elegir. Pero lo que pretendemos no es sólo hablar de la elección, sino de la elección racional. Si la condición para que podamos hablar de elección es que resulta imposible seguir un curso de acción sin desdeshar otros, la condición de que podamos hablar de elección racional es que no somos indiferentes ante la elección de varias alternativas. Cada uno de los cursos de acción alternativos entre los que hemos de elegir, así como las consecuencias que se siguen de cada uno de ellos, no sólo son distintos, sino que son relevantemente distintos, en el sentido de que los valoramos de un modo distinto: ver una película en el cine nos apetece más que quedarnos en casa, creemos que votar a tal partido político resultará más beneficioso para el país que votar a tal otro, la dedicación a la filosofía nos parece un proyecto de vida mejor que el de hacer dinero. Es decir, nuestro interés en que se produzcan determinados sucesos o en rea-

lizar determinadas acciones hace que podamos discriminar entre los distintos cursos de conducta alternativos de un modo relevante para la elección. Naturalmente, esto no quiere decir que *nunca* seamos indiferentes ante un conjunto de sucesos o de conductas. Pero el hecho de que no siempre lo seamos hace posible que podamos hablar de elección racional.

Podemos afirmar que "el concepto de conducta racional surge a partir del hecho empírico de que el comportamiento humano es en gran medida un comportamiento dirigido a fines" (Harsanyi, 77a, 627). Es decir, el hecho de que el hombre tenga determinados fines que intenta alcanzar mediante su conducta es lo que posibilita, en *primer* lugar, que no seamos indiferentes ante varias conductas alternativas, supuesto que cada una de estas conductas tendrá una consecuencia distinta de las demás y que no somos indiferentes ante las distintas consecuencias; y, en *segundo* lugar, que, en virtud de estas consecuencias realizadas mediante la conducta elegida, podamos valorar nuestras acciones; serán acertadas o desacertadas, buenas o malas, racionales o irracionales.

Antes de seguir adelante, conviene hacer algunas precisiones destinadas a evitar algunos malentendidos que pueden surgir en relación con lo dicho hasta ahora:

la *primera* precisión tiene que ver con la relación entre la conducta elegida y sus consecuencias. Supongamos que decido ir al cine un sábado por la noche a ver una película de gran éxito que acaban de estrenar en Madrid. Naturalmente, para conseguir el fin deseado (ver la película) tengo que hacer varias cosas: vestirme, coger un autobús, ir a la taquilla, pagar una entrada, etc. Y, consecuentemente, decido realizar estas acciones en vez de otras muchas posibles. Ahora bien, alcanzar la consecuencia deseada no depende solamente de que yo lleve a cabo todas estas acciones. Más bien al contrario, depende de otras muchas cosas, algunas de ellas de carácter muy general, tales como que funcionen los autobuses, y otras más específicas, como que haya entradas o que el autobús que tomo no tenga una avería. Sin embargo, el hecho de que en la realización de un efecto tomen parte muchas causas y que no todas ellas dependan de nosotros no rompe la conexión entre nuestras elecciones y acciones y la consecución de nuestros fines. Y esto sucede porque nuestras acciones son parte necesaria de la consecución de tales fines y porque al elegir una determinada acción tenemos en cuenta el resto de las circuns-

tancias necesarias para conseguir lo que deseamos, de tal modo que podamos tomar nuestras decisiones de manera que resulten adecuadas al resto de los factores.

La segunda precisión, o conjunto de precisiones, está relacionada con el carácter finalístico que hemos atribuido a la acción humana. Nuestra afirmación puede ser interpretada de dos modos, uno trivial y otro sustantivo. Todas las acciones tienen consecuencias. Por lo tanto, hay un sentido en el que siempre podemos aplicar el modelo medios/fines, pues siempre es posible hablar de las consecuencias de nuestras acciones como "fines" y de estas como "medios". Entendida en este sentido, nuestra afirmación de que la acción humana está dirigida a fines resulta trivial y, desde luego, no puede pretender surgir de ningún hecho empírico. Sin embargo, nuestra afirmación no sólo señala que nuestras acciones y sus consecuencias pueden ser reinterpretadas según el esquema finalístico, sino que sostiene la tesis sustantiva de que nuestra conducta está en gran medida *dirigida* a fines, es decir, que con nuestras acciones pretendemos conseguir determinados fines. Al contrario de lo que sucede con la interpretación trivial, la afirmación sustantiva no puede ser aplicada a todas las acciones por el sólo hecho de que tengan consecuencias, sino sólo a aquellas que están dirigidas a la consecución de fines de modo intencional. Pensemos por ejemplo en alguien que se muerde las uñas. Esta acción tiene como consecuencia unas uñas rotas. Podemos interpretar la acción como un medio para conseguir un fin, a saber el de tener las uñas rotas. Sin embargo, la acción no está dirigida a este fin. Y no sólo por el hecho de que el agente no desea tener las uñas rotas. La acción de morderse las uñas puede tener también otras consecuencias que el agente valora positivamente, como por ejemplo librarse de la ansiedad, sin que por ello pueda decirse que librarse de la ansiedad sea la finalidad de la acción de morderse las uñas. El agente no se dirige mediante esta acción a este fin, ni a ningún otro. Más bien se trata de una acción compulsiva¹.

¹ Esto no significa que el agente no realice la acción porque se encuentra en un estado de ansiedad, sino únicamente que no lo hace para librarse de ella. De hecho, en muchos casos de conducta compulsiva.

La afirmación sustantiva no se aplica sin más a toda acción humana. No se afirma que toda acción humana está dirigida a fines, sino que *en gran medida* lo está. Es decir, se afirma que una parte importante de la acción humana tiene un carácter finalístico. Y esta importancia no es sólo numérica. En efecto, se afirma también que es precisamente la acción humana dirigida a fines la que puede ser calificada de racional o irracional. Esto viene implicado directamente por nuestra afirmación anterior de que lo que hacía posible la elección racional era el hecho de que los distintos cursos de conducta alternativos y sus consecuencias no nos fueran indiferentes. Desde el momento en que podemos discriminar valorativamente entre varias alternativas podemos decidir nuestra conducta con vistas a la consecución de aquello que valoramos más, es decir, nuestra conducta se dirige hacia la consecución de un fin.

Queda pues establecido que la afirmación respecto al carácter finalístico de la acción humana debe ser entendida de un modo sustantivo. Sin embargo, la tesis sustantiva puede ser mal interpretada en dos aspectos fundamentales y, correspondientemente, pueden surgir dos objeciones contra ella. En primer lugar, puede alegarse que el hecho empírico en el que dice fundarse la tesis sustantiva es falso, pues no sucede que la acción humana esté por lo general dirigida a fines. Aunque es cierto que el hombre en ocasiones dirige su conducta a un fin, este tipo de conducta no es el único ni el más importante. Hay multitud de ocasiones en las que una acción no es realizada para conseguir un fin que se valora positivamente, sino que se realiza la acción porque ella misma es valorada positivamente. Es cierto que en ocasiones viajo en coche para llegar a algún sitio, pero en otras ocasiones viajo en coche sin que esta acción esté dirigida a la consecución de alguna otra cosa. Sin entrar a discutir si las primeras ocasiones son más o menos numerosas que las segundas, cosa por lo demás irrelevante, lo que sí es cierto es que, por un lado, no puede decirse que la acción humana pertenezca "en gran medida" al primer tipo ni, por otro lado, parece haber ninguna justificación para pretender que sólo las acciones dirigidas a un fin puede ser racionales.

aunque fumar pueda resultar tranquilizante, esto sólo es cierto para un fumador habitual, en el cual la nicotina tiene como efecto aliviar un estado de ansiedad generado por la carencia de nicotina, sustancia que de por sí no es precisamente un tranquilizante sino más bien todo lo contrario.

Esta objeción, sin embargo, se basa en una mala interpretación de "fin". "Medio" y "fin" son conceptos distintos que pertenecen a categorías contrarias, es decir, que no pueden ser aplicadas afirmativamente en el mismo sentido al mismo objeto. Esto significa que si algo es un medio entonces no puede ser, en tanto que medio, fin. En este sentido, puede decirse que un fin es necesariamente distinto del medio y "exterior" a él. Puede entonces pensarse que al decir que la acción humana está en gran medida dirigida a fines estamos queriendo decir que la acción humana es siempre, o casi siempre, un medio, lo cual es sin duda una afirmación dudosa. Sin embargo, que "medio" y "fin" pertenezcan a categorías contrarias no significa que algo que es un fin de una acción no pueda ser, al mismo tiempo y en otro sentido, un medio para otra, como por ejemplo cuando ir al cine es el fin de la acción de tomar un autobús y es también un medio de ver al taquillero. Tampoco significa que algo no pueda ser un medio en ocasiones y un fin en otras, como la acción de conducir es a veces un medio para otra cosa y a veces es un fin en sí misma. En efecto, en muchas ocasiones no realizamos acciones como medios para conseguir algo, sino que las realizamos por el valor que en sí mismas las atribuimos. Al decir que el comportamiento humano está en gran medida dirigido a fines no se está queriendo decir que las acciones no puedan ser fines, sino que el hombre tiene determinados fines que intenta conseguir mediante su conducta, siendo perfectamente posible que la realización de una determinada acción sea en sí misma un fin.

La otra objeción que puede plantearse a la afirmación finalística sustantiva es la siguiente. Hemos dicho que lo que hacía posible la elección racional era el hecho de la no indiferencia. Ahora bien, puede objetarse que de este hecho no se sigue sin más la verdad de la tesis finalista sustantiva. Dentro de la conducta humana atribuible a la no indiferencia se encuentran las acciones reactivas. Por ejemplo, la no indiferencia ante una sensación dolorosa hace que retiremos la mano del fuego. Desde luego, el fin de esta acción es evitar el dolor, pero parece que este tipo de finalidad de las acciones reactivas es sustancialmente distinta de la finalidad positiva a la que parece referirse la tesis sustantiva.

no sólo a acciones positivamente finalísticas, sino también a acciones reactivas. Sin embargo, no es cierto que la afirmación finalística sustantiva se refiera, o deba referirse, únicamente a las primeras. Tanto unas como otras cumplen el requisito definitorio de ser acciones dirigidas a la consecución de un fin de modo intencional. La confusión surge en parte porque al hablar de acciones reactivas suele pensarse en ejemplos que, como en el caso de la reacción de retirar la mano del fuego, son casi siempre actos reflejos y, por ello, no intencionalmente finalísticos. Pero esto no siempre sucede así. Supongamos que cojo una sartén del fuego sin darme cuenta de que el mango no es aislante. Puede que suelte la sartén de un modo reflejo, o puede que busque el sitio más cercano para colocarla sin dejarla caer al suelo. Que haga una u otra cosa depende de varios factores, siendo uno fundamental cuánto queme la sartén en cuestión. Ambas acciones son reactivas, y ambas surgen de la no indiferencia al dolor. Pero la cuestión fundamental para que estas acciones sean finalísticas en el sentido de la afirmación sustantiva no es si son reactivas o si son positivas, sino si son intencionalmente finalísticas o no, es decir, si son o no son un acto reflejo. Y los actos reflejos no son finalísticos en el sentido de la afirmación sustantiva porque no son intencionales y no porque sean reactivos. Es decir, no son finales por el mismo motivo por el que no lo son las compulsiones, a saber, porque aunque mediante ellos se alcance un determinado fin no están intencionalmente dirigidos a un fin. Por ello, aunque sean trivialmente finalísticos no lo son en el sentido de la tesis sustantiva, pero sin que esto tenga que ver necesariamente con el carácter reactivo o positivo de la acción.

La tercera precisión está relacionada con la relación que hemos establecido entre el carácter finalístico de la acción humana y la valoración de nuestras acciones. Hemos dicho que el hecho de que la conducta humana esté dirigida a fines hace posible que valoremos nuestras acciones calificándolas de acertadas o desacertadas, de racionales o irracionales.

Sin embargo, estos dos pares de calificativos no tienen un significado enteramente similar. Al calificar una acción de adecuada o inadecuada nos referiremos a su eficacia de hecho para producir el fin propuesto. Por ejemplo, si yo consigo ver la película que quiero ver, mis

acciones habrán sido las adecuadas, puesto que mediante ellas he conseguido el fin que me proponía, y habrán sido inadecuadas en caso contrario.

Por otro lado, al hablar de que nuestra conducta es racional o no lo es, haremos referencia más bien a si al elegir nuestra acción hemos tomado en consideración todos los hechos que pudieran hacer que nuestra acción fuera lo más eficaz posible dados nuestros fines, sea cual sea el resultado de nuestra acción.

Es fácil darse cuenta de que ambas cosas no tienen por que coincidir. Volvamos a nuestro ejemplo anterior. Si yo sé que la película que quiero ver acaba de ser estrenada y que está teniendo un gran éxito de público, puedo deducir que, con toda probabilidad, un sábado por la noche será muy difícil conseguir entradas. Por tanto, lo mejor que puedo hacer es salir de casa con tiempo suficiente para estar en la taquilla bastante antes de que empiece la sesión, o bien ir a sacar las entradas por la mañana. En la frase anterior, "lo mejor" quiere decir "lo más razonable", esto es, quiere decir que es la conducta racional a seguir dado que quiero ver esa película el sábado por la noche. Sin embargo, pueden ocurrir dos cosas: que, a pesar de seguir el curso de acción más razonable, no consiga ver la película, por ejemplo porque tenga un accidente al dirigirme al cine, o porque me roben la cartera o porque me encuentre con un amigo de la infancia que me entretenga para tomar una copa. Por otro lado, también puede suceder que, a pesar de haberme comportado de modo poco razonable presentándome en la puerta del cine diez minutos antes del pase, consiga ver la película, por ejemplo porque me encuentre por casualidad con alguien que quiere vender una entrada que le sobra. En el primero de estos dos casos la acción racional no resulta efectiva y por ello podemos decir que, en cierto sentido, no ha sido acertada. En el segundo caso la acción irracional tiene éxito, por lo que, también en cierto sentido, podemos decir que ha sido acertada.

A pesar de que la acción racional no siempre coincide con la acción acertada, podemos esperar que ambas características, e.d., ser racional y ser adecuada, coincidan en la mayoría de los casos. Esto, naturalmente, no es ninguna sorpresa, ya que definimos la acción racional

aunque la acción racional no sea infalible sí es la que tiene más probabilidades de éxito. De hecho, al considerar en que circunstancias la acción irracional habría resultado exitosa hemos tenido que utilizar expresiones como "a pesar de" y "por casualidad", de igual modo que al explicar lo que hubiera podido hacer fracasar a la acción racional hemos tenido que acudir a accidentes. Aunque esto no sea una sorpresa, conviene tener presente que esto no sucede porque de hecho la acción racional suele tener buenos resultados, sino que más bien la acción racional *es* la acción racional *precisamente* porque es la acción con más garantías de éxito.

Existen buenos motivos para que nos ocupemos de la conducta racional y no de la conducta exitosa. El primero es que estamos sin duda interesados en saber qué acciones nos llevarán a la consecución de nuestros fines y la acción racional es por definición la que tiene más garantías de éxito. El segundo motivo está estrechamente emparentado con el primero, y es que si bien podemos saber *antes* de actuar cuál es la acción racional en cada caso, sólo *después* de realizar la acción podemos saber si ha tenido o no ha tenido éxito. Y, naturalmente, esto no sólo significa que *no puede* haber una teoría sobre la acción acertada (salvo en tanto que esta coincide con la acción racional), sino que no estamos en absoluto interesados en saber lo mejor que podemos hacer cuando ya está hecho, y que estamos extremadamente interesados en saberlo antes, *e.d.*, cuando podemos aplicar lo que sabemos.

Hemos visto que podemos valorar las elecciones y acciones en relación con los fines que queremos alcanzar. Estos fines nos proporcionan el criterio de la acción racional, de tal modo que podemos definir en principio una acción racional como aquella que consiste, o bien en la puesta en práctica de los medios más adecuados para la consecución de un fin dado, o bien en la realización de la acción más valorada en sí misma².

² Esta definición tentativa de "acción racional" pretende recoger tanto el caso de las acciones que son medios para fines como el de aquellas acciones que constituyen fines en sí mismas. De ahora en adelante, aun cuando por motivos de brevedad sólo se mencione el primer caso, habrá de entenderse que, tal y como se ha aclarado en la discusión de la página 4, el caso de las acciones que son fines en sí mismas también queda recogido en la definición.

Los fines que los agentes se proponen son múltiples. Es un hecho empírico incuestionable el que los hombres dirigen su acción a conseguir una multiplicidad de objetivos diversos. Esta diversidad se encuentra en varios niveles. En el nivel interpersonal, distintos agentes se proponen diversos fines. Por otro lado, en el nivel intrapersonal, un mismo agente se propone distintos objetivos en distintos momentos de su vida. Esta multiplicidad es indiscutible en el nivel de fines intermedios, es decir, aquellos que, siendo fines particulares de una acción particular, no son queridos por sí mismos sino en virtud de alguna otra cosa con respecto a la cual son medios. Sin embargo, puede pensarse que aunque esta multiplicidad de fines intermedios es irreductible, todos ellos adquieren su valor en virtud de su carácter de medios con respecto a un fin último que se supone único. En cualquier caso, para nuestro propósito basta con afirmar que esta multiplicidad se da en el nivel de fines intermedios, pues estamos interesados en la elección racional entre acciones alternativas en virtud de nuestros fines, siendo indiferente si estos son o no fines intermedios*.

Para recoger esta diversidad empírica de propósitos, es habitual hablar de "deseos" o "preferencias" como nombre genérico para agrupar todo aquello que los distintos hombres se proponen como fin. El término "deseo" indica que el fin de una acción determinada será aquello que el agente se propone conseguir mediante ella. Sin embargo, el término es equívoco por dos motivos. En primer lugar, porque está asociado a un tipo determinado de razón por la que el agente puede proponerse conseguir algo como distinto a otro tipo de razones. Así por ejemplo, yo puedo proponerme ver una película porque me apetece, pero también porque debo preparar una crítica sobre ella. Puesto que el término "deseo" es habitualmente empleado en casos como el primero, puede inducir a confusión su uso para cubrir todos los casos.

En segundo lugar, el término "deseo" sólo indica el hecho de que el fin de una acción es algo querido por el agente. Esto podría llevar a formular la relación entre acción racional y fines diciendo que una acción racional es aquella que emplea los medios más efectivos posibles para la satisfacción de los deseos del agente. Sin embargo, esta iden-

tificación o definición de acción racional no es aceptable, pues ignora el hecho importante de que un agente tiene muchos deseos y que estos son de tal modo que por lo general no es posible que todos resulten satisfechas. Por ejemplo, yo tengo un deseo de ver una película el sábado por la noche, pero también tengo otro, madrugar el domingo para ir a pasear. Supongamos para simplificar que ambos deseos no pueden cumplirse al mismo tiempo, lo cual es por otra parte un supuesto muy razonable, dado que el ser humano en general necesita dormir y yo en particular necesito dormir bastante. Precisamente porque se dan estos casos es por lo que necesito elegir y necesito una teoría sobre la elección racional. Si sólo tuviéramos un deseo a un tiempo, la anterior definición de acción racional sería suficiente y la elección inmediata.

Ahora bien, si tuviéramos deseos distintas e incompatibles y los deseáramos en el mismo grado, aunque tendríamos un problema de elección no tendríamos una solución. Sin embargo, esto no sucede. No todas las cosas que queremos las queremos con la misma intensidad. Pensemos por ejemplo en un ama de casa que va a hacer la compra. Supongamos para simplificar que en el supermercado sólo hay cuatro artículos disponibles, a saber, pan, cerveza, té y azúcar. Ella desearía comprar todas estas cosas pero como no tiene suficiente dinero se ve obligada a elegir. Elegir una de estas cosas significa renunciar a las demás. Es de esperar que el ama de casa decida comprar aquello que desea o necesita en mayor medida. Por ejemplo, puede que sea té lo que prefiere. El término "preferencia" señala este hecho, a saber, que entre todos las posibles alternativas que podemos proponernos como fines de nuestra acción, incluso cuando todas o varias de ellas puedan resultarnos en alguna medida deseables, podemos establecer entre ellas un orden de preferencia.

Tenemos pues una lista de alternativas que podemos ordenar jerárquicamente según un orden de preferencia. Es decir, están ordenadas según el grado en que deseamos conseguir las. Habitualmente, se llama "utilidad" a ese grado distinto de satisfacción que asociamos con la consecución de cada una de las posibles alternativas. Así, si mi preferencia por ver la película es más fuerte que mi preferencia por ir a pasear, diremos que la

utilidad de ver la película es mayor. En este sentido, podemos decir que la utilidad es la medida de nuestras preferencias*.

Podemos ahora definir la acción racional de una manera más precisa y adecuada utilizando el concepto de utilidad. Diremos que una acción racional es la que maximiza esa medida, e.d., la que maximiza la utilidad del agente. Dicho de otro modo, la acción racional será aquella que consiste en la puesta en práctica de los mejores medios posibles para conseguir, de entre todas las alternativas posibles, aquella que preferimos en mayor medida. Una acción racional será por tanto aquella que tiene como resultado conseguir el mejor resultado posible, e.d., el máximo de utilidad posible para el agente.

* Podría pensarse que es necesario añadir "la utilidad para mí". Sin embargo, este añadido es completamente innecesario, pues sólo estamos considerando la utilidad como una medida sobre preferencias individuales, de tal modo que se sobreentiende que es utilidad para mí desde el momento en que estamos estableciendo una medida para mis preferencias. Por tanto, en este momento nos referimos a la utilidad como una medida puramente subjetiva.

CAPITULO II

CONDICIONES DE RACIONALIDAD

En el capítulo I vimos que la acción racional sólo es posible si el agente puede establecer un cierto orden entre las distintas alternativas que se le ofrecen y llamamos a este orden "orden de preferencias". Ahora bien, esto no define completamente el concepto de elección racional. En *primer* lugar, necesitamos un análisis de las distintas relaciones formales que deben establecerse entre estas alternativas para que estas puedan quedar ordenadas de tal modo que la acción racional sea posible. En *segundo* lugar, debemos preguntarnos si cualquier conjunto de alternativas cuyos elementos guarden entre sí estas relaciones hace posible la elección racional o si, por el contrario, debemos establecer también determinadas condiciones materiales que las preferencias deben cumplir para poder hablar de acción racional. En este capítulo nos ocuparemos de ambas cuestiones en ese orden.

1. Condiciones formales de racionalidad.

Hemos visto que la conducta racional se identifica con la maximización de la utilidad. Ahora bien, para que sea posible emplear la utilidad como una medida sobre preferencias es necesario que estas cumplan con una serie de requisitos.

Estas condiciones, que deben cumplir las preferencias de una persona para que sea posible hablar de elección racional entre ellas, son de carácter formal. Es decir, son condiciones que deben cumplir, no las preferencias tomadas aisladamente, sino los conjuntos de preferencias de una persona, y hacen referencia a las relaciones de las preferencias entre sí.

Es habitual referirse a estas condiciones diciendo que el conjunto de preferencias de un individuo debe formar una *ordenación* para que la elección racional sea posible. Una ordenación es una relación de orden entre los miembros de un conjunto, en nuestro caso entre las preferencias de un individuo.

Sin embargo, como señala Sen (Sen, 76), la terminología no está unificada, y distintos autores exigen distintas condiciones para que una relación de orden constituya una ordenación. Concretamente, según algunos autores una ordenación debe cumplir las propiedades de transitividad, reflexividad y completud, mientras que en otros autores esta última propiedad se substituye por la de antisimetría. Veamos las definiciones de estas propiedades:

Sea R la relación binaria que queremos definir

Sea S el conjunto sobre el que se define la relación R y

Sean x, y, z , miembros de S .

entonces R es reflexiva cuando $\forall x \in S : xRx$

es transitiva cuando $\forall x/y/z \in S : (xRy \ \& \ yRz) \longrightarrow xRz$

es antisimétrica cuando $\forall x/y \in S : (xRy \ \& \ yRx) \longrightarrow x=y$

es completa cuando $\forall x/y \in S : (x \neq y) \longrightarrow (xRy \vee yRx)$ ¹

¹ Los símbolos lógicos utilizados tratan de seguir las convenciones habituales. Así ' \forall ', ' \vee ', ' \neq ' y

Tanto la reflexividad como la transitividad son generalmente consideradas como condiciones que una relación de orden debe cumplir para ser una ordenación. Sin embargo, no hay acuerdo respecto a las otras dos propiedades. Nosotros seguiremos a autores como Arrow y Sen y entenderemos que una ordenación debe ser completa y que esta condición junto con la reflexividad y la transitividad caracterizan una ordenación sin tener en cuenta la antisimetría.

El motivo por el que no tendremos en cuenta la antisimetría se fundamenta en el tipo de relación en el que estamos interesados, es decir, la preferencia. Partimos de la base de que, para que sea posible la elección racional, es necesario que las preferencias de un individuo guarden entre sí cierto tipo de relación, a saber, una relación que permita compararlas mediante la utilización de una medida, a la que hemos llamado utilidad. Ahora bien, lo que necesitamos no es una relación de preferencia estricta, sino una que incluya también la noción de indiferencia. Es decir, la relación R que nos interesa definir no es la relación "ser preferido a" sino la relación "ser preferido o indiferente a". A esta segunda relación la llamaremos simplemente de preferencia (R), mientras que nos referiremos a la primera como relación de preferencia estricta (P).

Veamos la razón por la que caracterizamos de este modo a la relación que nos interesa definir. En primer lugar, la relación de preferencia estricta parece demasiado fuerte e innecesariamente restrictiva. Sin duda alguna, si un individuo fuera indiferente a todos los miembros de su conjunto de preferencias nos resultaría imposible mantener la posibilidad de una elección racional entre ellas. Necesitamos que las preferencias estén ordenadas jerárquicamente. Pero no parece necesario prohibir que pueda haber dos preferencias compartiendo un mismo puesto en la jerarquía. Además, esta exigencia sería poco realista.

La antisimetría es una relación que sólo pueden cumplir las relaciones estrictas, e.d., las relaciones que no admiten empates. Por ejemplo, la relación "más alto que" es antisimétrica, pues si x es más alto que y y y es más alto que x , entonces necesariamente x y y son el mismo individuo. Sin embargo, la relación "al menos tan alto como" no es anti-

*A se usan con el significado habitual.

simétrica, pues no es en absoluto necesario que dos individuos sean el mismo individuo para ser de la misma estatura².

Por tanto, la relación que nos ocupa no cumple la antisimetría, sin que ello suponga un impedimento para poder establecer entre las preferencias un orden jerárquico que nos permita utilizar una medida para ellas³.

Una vez que hemos fijado la terminología y que hemos explicado por qué razón no exigimos que la relación entre preferencias sea antisimétrica, podemos pasar a analizar las condiciones que exigimos y la razón para imponer tales condiciones.

Respecto a la *reflexividad* como condición de racionalidad de las preferencias podemos decir que no suscita debate, ya que es generalmente considerada casi como una condición de sentido común (Sen 76). En efecto, lo único que esta condición exige es que cada elemento del conjunto de preferencias sea considerado al menos tan bueno como él mismo⁴.

La *completud* exige que cada uno de los elementos del conjunto este relacionada con todos los demás. En nuestro caso esto se concreta así: siendo la relación que nos ocupa la de "ser preferido o indiferente a", para todo par de alternativas x e y debe darse o bien el par ordenado (x,y) o el par (y,x) o ambos. Si sólo apareciese uno de los pares, la relación entre ambos elementos sería la de preferencia estricta y en el caso de que aparecieran ambos la relación sería de indiferencia.

Hemos dicho que la elección guarda una estrecha conexión con las preferencias. Esta conexión consiste en que tenemos motivos para elegir una cosa y no otra sólo bajo el supuesto de que tenemos preferencias. Por esto decimos que una elección es racional si es la que posibilita la rea-

² La única excepción la constituyen las relaciones numéricas, como "mayor o igual que" y "menor o igual que", que sí cumplen con la propiedad de la antisimetría, ya que si x es igual que y , entonces necesariamente $x = y$ y son el mismo número, ya que precisamente ser mayor o menor que otros es el único criterio de identidad para los números.

³ Algunos autores, por ejemplo Gauthier (86) exigen como nosotros que la relación entre los miembros de un conjunto de preferencias sea reflexiva, transitiva y completa, pero, sin embargo, no llaman a esta relación ordenación sino "ordenación débil". Nuestra elección respecto al nombre de "ordenación" obedece por tanto a que por lo general este es el término utilizado. Por lo demás, la elección del nombre es indiferente siempre y cuando quede claro a que propiedades estamos haciendo referencia.

lización de las preferencias del agente³. Esto significa que las alternativas entre las cuales se realiza la elección tienen que ser comparables en términos de preferencias. Y esto es precisamente lo que exige la completud, e.d., que los miembros de cualquier par de alternativas puedan compararse.

Podemos ver hasta qué punto la completud es una condición de racionalidad mediante un ejemplo. Supongamos que puedo elegir entre las siguientes alternativas: a-ir al cine, b-quedarme en casa y c-ir al teatro. Realizaré la elección según mis preferencias. Supongamos que estas alternativas se estructuran según la relación R y que R es completa y reflexiva. Por ejemplo,

sea $R \{ (a,a), (b,b), (c,c), (a,b), (a,c), (b,c), (c,b) \}$

La alternativa a es estrictamente preferida a las otras dos, y las alternativas b y c son indiferentes entre sí. Puesto que las tres alternativas son comparables entre sí, podemos elegir entre ellas. En nuestro ejemplo, la elección recaería en la alternativa a.

Supongamos ahora que R no es completa. Por ejemplo, digamos que R se define mediante las pares $\{ (a,a), (b,b), (c,c), (a,b) \}$. En este caso, el elemento c no está en relación con ninguno de los otros dos. Es decir, cuando nos enfrentamos con a y c, por ejemplo, no podemos decir cuál de ellas es preferida, ni tampoco que ambas son indiferentes. Sencillamente, no podemos decir nada de ellas respecto a la preferencia. ¿Podríamos elegir racionalmente en este caso?⁴. Que R sea completa es un requisito indispensable para que se puedan estructurar las preferencias en un orden jerárquico, pues un orden jerárquico supone que los elementos del conjunto están ordenados respecto a una relación. Ahora bien, si la relación no nos dice nada respecto a algún elemento del conjunto, entonces ese elemento no puede ser colocado en ningún puesto de la jerarquía. Los elemen-

³ Esto es condición necesaria aunque, como veremos más adelante, no suficiente.

⁴ Gauthier defiende que si bien la completud es una condición necesaria para la elección racional, no lo es para la posibilidad de elección (Gauthier 86,p.39). Es decir, aunque sea imposible elegir racionalmente entre alternativas inconmensurables, es posible elegir entre ellas.

El argumento de Gauthier es bastante discutible. La discusión es interesante en tanto que clarifica los conceptos de elección y preferencia. Sin embargo, he preferido no tratarlo aquí, pues podría desviarnos bastante y en este momento nos basta con mostrar que la completud es necesaria para la elección racional.

tos no relacionados por R serían incommensurables entre sí. Y, como quedó explicado anteriormente, esto hace imposible una elección racional entre ellos.

La importancia y el significado del requisito de completud pueden clarificarse considerando una crítica que me parece significativa (Sen y Williams 82).

La crítica tiene dos pasos:

1. Sen y Williams argumentan que lo que hace irracionales ciertas conductas, como por ejemplo la representada por el asno de Buridán, no es la incapacidad para ordenar dos montones de heno, sino la negativa a elegir hasta no estar seguro de cual de los dos es el mejor, o al menos tan bueno como el otro.

2. A continuación se presenta una alternativa: puede haber elección racional basada en un orden incompleto, y la condición requerida para hacer racional esta elección es que la alternativa elegida no sea inferior a la otra. En el caso del asno, este se hubiera comportado racionalmente si hubiera elegido cualquiera de los dos montones.

El ejemplo propuesto no es muy afortunado. En primer lugar, porque no se entiende muy bien cómo puede ser que dos montones de heno sean incommensurables, de modo que no pueda decirse si uno es mejor que el otro o si los dos son iguales. Y, en segundo lugar, porque lo que le sucedía al asno era que no podía decidirse porque ambos montones le parecían igual de buenos. Pero esto no es un caso de relación incompleta, sino un caso de empate, esto es, de indiferencia debida en este caso a la indiscernibilidad de las opciones.

En los casos de empate puede ser o no irracional no elegir. Esto depende de si el empate es entre alternativas igual de buenas o igual de malas: si yo puedo elegir entre un helado de fresa y uno de chocolate, parece una elección irracional que me quede sin ninguno, pero si tengo que elegir entre sesos y criadillas y puedo elegir también no comer ninguna de las dos cosas (e.d., si esta elección no supone que me voy a morir de hambre) no es irracional no decidirme por ninguna de las dos.

Ahora bien, en el caso del asno es racional elegir, tanto porque ambos son igual de buenos como porque la alternativa restante es morir

dos. Pero esto habría sido lo racional porque se trata de un caso de empate. La única condición en estos casos es que la alternativa elegida no sea inferior a la otra. Pero si el caso hubiera sido de auténtica incommensurabilidad no tendría sentido hablar de "inferior", pues de dos cosas que no son comparables no puede decirse que ninguna sea inferior a la otra. Y si se puede ser inferior, entonces se puede ser superior o igual.

Este tipo de críticas surgen de la confusión entre indiferencia y falta de completud, confusión contra la que el propio Sen nos previene (Sen, 76). Para evitar la posibilidad de esta confusión expondremos la diferencia lógica entre ambas: en caso de empate se da xRy y también yRx . En caso de incompletud no se da ninguno de los dos.

La última condición que caracteriza a una ordenación es la *transitividad*. Esta propiedad se exige no sólo en la ordenación, sino en todas las relaciones de orden. La idea intuitiva subyacente es que las relaciones de orden establecen, precisamente, un orden jerárquico, e.d., una relación en la que los elementos del conjunto aparecen de mayor a menor. En el caso que nos ocupa, los elementos aparecerán ordenados de más a menos preferido. Por ejemplo, si yo puedo elegir entre ir al cine, al teatro o quedarme en casa, estas alternativas deben estar jerarquizadas según mis preferencias para que pueda hablarse de elección racional. Una posible jerarquización de estas alternativas sería, por ejemplo, ir al cine, quedarme en casa, ir al teatro.

La transitividad exige que si ir al cine ocupa un puesto superior en la lista a quedarme en casa, y si quedarme en casa ocupa un puesto superior a ir al teatro, entonces ir al cine ocupa un puesto superior a ir al teatro. Ahora bien, parece que esto no exige otra cosa salvo que los miembros del conjunto estén jerarquizados. Para ver esto, supongamos que el conjunto de preferencias no es transitivo. Sea a ir al cine, b quedarme en casa y c ir al teatro. Una violación de la transitividad supondría que a está por encima de b, b por encima de c y c por encima de a. Esto quiere decir que nuestra lista no establece un orden jerárquico, sino más bien un círculo de preferencias, donde la primera es preferida a la segunda, la segunda a la tercera y la tercera a la primera (suponiendo que en este caso aun tuviera sentido hablar de "primera" etc.). Evidente-

mente, la existencia de estos círculos impediría una elección racional entre las alternativas.⁷

Con esto hemos finalizado nuestro análisis de las propiedades de una ordenación. Sin embargo, se admite generalmente que para que haya elección racional no basta con que la elección se realice a partir de un conjunto de preferencias que formen una ordenación. Aparte de estas condiciones acerca de la estructuración de un conjunto de preferencias, se cree que es necesario establecer también una serie de condiciones sobre las preferencias tomadas individualmente. Al estudio de estas condiciones estará dedicada la segunda parte de este capítulo.

⁷ Se acepta generalmente que la posibilidad de preferencias cíclicas debe ser excluida del ámbito de la elección racional. Sin embargo, algunos autores consideran que no es necesario exigir transitividad para evitar estos círculos, sino que basta con exigir la condición más débil de aciclicidad. Para una discusión detallada de las condiciones de aciclicidad y transitividad, así como de las razones que pueden aducirse a favor de la exigencia de transitividad y no sólo de aciclicidad, ver Apéndice 2.

2. Condiciones materiales de racionalidad.

En la primera parte de este capítulo hemos visto cuál debe ser la estructura de un conjunto de preferencias para que la elección racional sea posible. Son, tal y como hemos visto, condiciones necesarias para la elección racional. Esto, sin embargo, no significa que sean también condiciones suficientes. Por el contrario, parece que un conjunto de preferencias puede ser tal que no pueda seguirse de él una elección racional, y esto no en virtud de la estructuración del conjunto sino a causa del carácter de sus elementos. Es decir, parece que no sólo pueden distinguirse conjuntos racionales y conjuntos irracionales de preferencias, sino que también puede establecerse una distinción entre preferencias racionales y preferencias irracionales.

Una ordenación es una estructura. Sean cuales sean las preferencias de un individuo, estas pueden estar relacionadas entre sí de tal modo que el conjunto tenga la forma de una ordenación. Ahora bien, si podemos distinguir entre preferencias racionales e irracionales, entonces no basta con que una elección se siga de una ordenación para que sea una elección racional, sino que será preciso que la ordenación se establezca entre las preferencias racionales de un individuo.

Las condiciones que una preferencia debe cumplir para ser una preferencia racional pueden dividirse en dos grupos. En primer lugar, están las condiciones impuestas no sobre el contenido concreto de las preferencias, sino sobre el modo en que estas surgen y se mantienen. Básicamente, estas condiciones exigen que las preferencias se hayan formado cuando el agente disponía de toda la información relevante, habiendo considerado la cuestión cuidadosamente y encontrándose en un estado de ánimo adecuado en el momento de realizar la reflexión.

La primera de estas condiciones se establece atendiendo a la conexión existente entre las preferencias que formamos sobre las cosas y el conocimiento que tenemos acerca de ellas. La información relevante se define en función del papel que tal información ejerce en la alteración de una preferencia, de modo que identifiquemos la información relevante con aquella cuya posesión haría que nuestras preferencias cambiaran. Por ejemplo, yo puedo tener formada una preferencia acerca de una película,

de modo que deseo verla, bien porque me gusta el director o los actores. Pero es posible que si conociera el argumento mi preferencia se invertiría, por ejemplo porque la película trate de un tema que me desagrada en extremo.

La segunda condición hace referencia a la necesidad de calcular los costos de una acción. Por ejemplo, mi preferencia por ir al cine puede haberse formado sin considerar que eso supone trasnochar demasiado. Si hubiera pensado en ese inconveniente, mi preferencia se habría alterado.

La tercera condición está relacionada con la segunda, en el sentido de que determinados estados de ánimo pueden afectar la reflexión. Por ejemplo, si he bebido un par de copas tengo cierta tendencia a pensar más en las ventajas que en los inconvenientes, lo cual afecta a la formación de mis preferencias en esos momentos. El estado de ánimo puede afectar a la formación de preferencias en otro sentido. Supongamos por ejemplo que durante una crisis de ansiedad alguien me obliga a beber tila. Puede suceder que a partir de ese momento yo adquiera una considerable aversión hacia las infusiones. Esta aversión es irracional en el sentido de que puede ser cambiada mediante una reflexión serena acerca de las virtudes de las infusiones y una comprensión del motivo que me ha llevado a formar mi antipatía.

Cumplir estas condiciones significa que nuestras preferencias, o bien se han formado en unas condiciones óptimas en cuanto a la información y al procesamiento de esta información, o bien que, a pesar de haberse formado en condiciones subóptimas, se mantienen una vez que han sido expuestas a una crítica a la luz de los hechos. En este sentido, podemos admitir la definición de Brandt² que identifica una preferencia racional como aquella que sigue manteniéndose después de un proceso de terapia cognitiva, donde por "terapia cognitiva" se entiende el proceso de confrontar los deseos con la información relevante mediante la representación insistente de esa información de un modo vivido y en el momento adecuado. Alternativamente, podemos decir que una preferencia es racio-

nal si se ha formado a partir de, o puede sobrevivir a, un proceso de deliberación ideal⁹.

El segundo grupo de condiciones se imponen sobre el contenido concreto de las preferencias. En este grupo se incluyen por un lado las condiciones sobre las preferencias individuales y por otro las condiciones sobre el conjunto de las preferencias de un agente respecto a su inclusión o no de ciertas preferencias.

En primer lugar, una preferencia puede ser intrínsecamente irracional¹⁰. Diremos que una preferencia es intrínsecamente irracional cuando es una preferencia hacia lo peor¹¹. Esta definición necesita muchas aclaraciones para poder ser bien entendida. En ocasiones, un agente tiene una preferencia por lo peor producida por un mal conocimiento de lo preferido o por un estado anímico perturbado. Es decir, en ocasiones la preferencia por lo peor es una preferencia que no podría sobrevivir a un proceso de deliberación ideal. Supongamos una persona (a quien nos referiremos como "el masoquista") que siempre prefiere el mayor de dos males. Es posible que su preferencia se deba a su creencia de que soportando los

⁹ Exigir que las preferencias se hayan formado de hecho cuando el agente conocía todos los datos relevantes, estaba pensando con claridad y no estaba sometido a influencias distorsionantes sería innecesariamente restrictivo. Dejaría fuera de consideración, por ejemplo, la mayoría de las preferencias originadas en la infancia. Además, relacionaría de un modo ilegítimo el origen de las preferencias con su racionalidad. Si, por ejemplo, el preferencia por la novela histórica se debe a la influencia de el padre en una etapa de el vida en la que yo carecía de los elementos necesarios para juzgar por mí misma, tendría que considerarla irracional simplemente por este hecho, incluso si con el paso del tiempo y una vez que yo tengo el conocimiento necesario mi preferencia se mantiene. Sin embargo, parece claro que lo fundamental para juzgar acerca de la racionalidad de una preferencia es que esta sobreviva a la confrontación con los hechos cuando el agente se encuentra en el estado de ánimo adecuado, con independencia de su origen.

¹⁰ Sobre el tratamiento de este tipo de irracionalidad de las preferencias, ver Parfit (86), pp.121 y ss.

¹¹ No debe confundirse este tipo de irracionalidad con el problema de la debilidad de la voluntad. Este último surge no cuando el agente prefiere lo peor (lo que él considera peor), sino cuando prefiere lo mejor (lo que él considera lo mejor) y sin embargo elige el curso de acción cuyo resultado es el peor. Precisamente por eso es un problema de debilidad de la voluntad. Algunos autores han negado la posibilidad de este fenómeno, atribuyendo los casos aparentes de debilidad de la voluntad a una preferencia, al menos asentada, por la peor alternativa. Sin embargo, como veremos más adelante, esto es difícilmente defendible. En cualquier caso, se trata de dos fenómenos distintos. En el caso que nos ocupa ahora, se trata de una irracionalidad debida al contenido de nuestras preferencias. En el caso de la debilidad de la voluntad, se trataría más bien de una conducta cuyo carácter irracional es debido a que no se sigue de una ordenación establecida sobre un conjunto de preferencias racionales.

mayores sufrimientos alejard de sí la mala suerte o a que de este modo atenda un sentimiento de culpa que él cree justificado. Es de suponer que el masoquista, sometido a una terapia adecuada, cambiará sus preferencias. En estos casos la irracionalidad de la preferencia por lo peor no es independiente del hecho de que no pueda sobrevivir a un proceso de deliberación ideal. Por tanto, dejaremos a parte estos casos.

En otros casos, un agente podría ofrecer razones que justificaran su preferencia por lo peor. Por ejemplo, supongamos que el masoquista puede justificar su preferencia diciendo que lo realmente importante para él es la posesión de una fortaleza de carácter inmune al dolor. De este modo, el masoquista supone que la elección del mayor entre dos males le resulta beneficioso¹². En este caso, la preferencia por lo peor no es irracional. Tanto en este caso como en el anterior, la preferencia del masoquista no es una preferencia por lo peor en sí mismo, sino por lo peor en tanto que medio para conseguir algo que consideramos valioso. La diferencia entre ambos casos es que en este último nuestra preferencia por lo peor sobrevive al proceso de deliberación ideal.

La afirmación acerca de la irracionalidad de preferir lo peor puede ser mal interpretada de otro modo. Según esta interpretación, podría afirmarse que un agente prefiere lo peor a pesar de que el agente en cuestión considera que su preferencia no es de lo peor sino de lo mejor. Supongamos que el masoquista, al ser interrogado respecto a por qué prefiere siempre lo peor, responde que esa acusación es falsa. "Es cierto", diría él, "que siempre prefiero el mayor de dos dolores. Pero esto no es preferir lo peor, sino lo mejor. Y sé que es lo mejor porque es lo más placentero." Supongamos además que el masoquista ha pasado por un proceso de deliberación ideal y que no elige el dolor por otra cosa sino por sí mismo (puede dudarse que los masoquistas reales sigan prefiriendo el dolor en este caso, pero esto es irrelevante aquí). En este caso, podemos decir que nuestro masoquista tiene unos gustos un tanto exóticos, pero esto no justifica la afirmación de que sean irracionales. Aunque este caso pueda no ser el más habitual en este tipo de confusión, si resulta

¹² Aunque este ejemplo es extremo, todos en algún momento tenemos una preferencia por lo peor que podemos justificar. Por ejemplo, podemos elegir el más doloroso de dos tratamientos si pensamos que así nuestra curación será más rápida o más efectiva.

ser paradigmático e ilustra un tipo de preferencias que no pueden ser irracionales en este sentido. Este grupo está formado por las preferencias ligados a placeres y dolores físicos

"Me gustan las duchas frías. Otros las detestan. Ninguno de estos deseos es irracional. Si yo quiero comer algo porque me gusta su sabor, este deseo no puede ser irracional. No es irracional incluso si lo que a mí me gusta les disgusta a todos los demás. Consideremos a continuación las experiencias que encontramos desagradables. Mucha gente tiene un fuerte deseo de no oír el chirrido de la tiza. Este deseo es extraño, puesto que a esa gente no le importa oír otros chirridos que son similares en timbre y tono. Pero este deseo no es irracional. Afirmaciones similares son aplicables a lo que encontramos doloroso."¹³

En lo que respecta a este tipo de preferencias, a las que en adelante nos referiremos con el nombre de "*preferencias primarias*", el propio individuo es el único que pueda juzgar. No es posible encontrar ningún argumento para apoyar la opinión de que algo de este tipo debe o no debe ser deseado. Como suele decirse, es simplemente una cuestión de gustos. Por ello, si alguien elige lo que nosotros consideramos peor, mientras que el lo considera mejor, su preferencia no es irracional.

Tras hacer estas aclaraciones, podemos interpretar correctamente lo que quiere decirse al afirmar que una preferencia es intrínsecamente irracional si es una preferencia de lo peor. Una preferencia es intrínsecamente irracional si, tras haber realizado un proceso de deliberación ideal, el agente prefiere lo que *él mismo considera peor sin ninguna razón que justifique esta preferencia* relacionándolo como medio con la consecución de un fin deseado como bueno.

Puede pensarse que ninguna preferencia satisface esta definición y que, por tanto, ninguna preferencia es irracional en este sentido. Sin embargo existen tales casos. Algunos son extraordinariamente habituales. Imaginemos a alguien que se come las uñas. La mayoría de la gente que hace eso considera que es peor comerse las uñas que no hacerlo. Tampoco pueden justificar su deseo de comerse las uñas relacionándolo con alguna otra cosa. Y en la mayoría de los casos este deseo sobrevive a un proceso de deliberación ideal. Desde luego, casi con seguridad este deseo ha surgido en condiciones que no son ni mucho menos de deliberación ideal, pero

¹³ Parfit (86), p.123.

una vez formado este deseo con gran frecuencia resulta inmune a la reflexión. La gente que se come las uñas por lo general es consciente de todos los datos relevantes y no se encuentra psicológicamente alterada, a menos que hagamos equivalente por definición comerse las uñas y estar psicológicamente perturbado. Otros casos son menos habituales pero mucho más importantes, como el caso citado por Parfit¹⁴ del deseo que algunas personas sienten de saltar cuando se encuentran al borde de un abismo.

La mayoría de las fobias generan preferencias irracionales en este sentido. Algunos casos se resuelven mediante el sometimiento del paciente a algún tipo de terapia cognitiva. Cuando esto sucede, la irracionalidad de la preferencia puede atribuirse al hecho de que no pueden sobrevivir a un proceso de deliberación ideal. Pero en muchísimos otros casos el deseo sobrevive a la terapia cognitiva. El agente no apoya sus preferencias en ninguna idea confusa, equivocada o en algún otro modo deficiente. Es más, sabe que su preferencia no obedece a ninguno de estos motivos. Es entonces cuando son propiamente compulsiones. Estas constituyen ejemplos legítimos de deseos y preferencias intrínsecamente irracionales¹⁵. De nuevo es necesario señalar que estas preferencias no obedecen a que el agente este psicológicamente perturbado, salvo en el sentido trivial en el que la perturbación consiste en tener tales deseos. Pero no en el sentido de que estos deseos se tengan porque se está psicológicamente perturbado, pudiendo definirse esta perturbación de modo independiente. De hecho, los agentes que tienen algún deseo irracional en este sentido desean verse libre de ellos y, tras la terapia cognitiva, desean voluntariamente someterse a algún otro tipo de tratamiento de carácter conductista, al menos en los casos en los que la terapia no es demasiado costosa y la compulsión es suficientemente importante.

¹⁴ Parfit (198), p.122.

¹⁵ Todos estos ejemplos pueden ser analizados como casos de debilidad de la voluntad. De hecho pueden serlo y es incluso posible que lo sean en muchos casos. Sin embargo, esto no sucede necesariamente. Es posible que alguien se acuerda las uñas porque prefiere hacerlo a pensar de creer que sería mejor no comérselas. Podría dejar de hacerlo si quisiera, pero no quiere. En el análisis de casos reales, es difícil decidir cual de estas dos cosas sucede. El psiquiatra puede pensar que el paciente cree que podría dejar de hacerlo pero que en realidad no podría. Probablemente tenga razón, pero no interesa aquí la discusión psiquiátrica. Lo que interesa es que son casos distintos (ver nota 9) y que el caso de irracionalidad de preferir lo peor es posible.

Puede dudarse de que, descontando algunos casos como los mencionados, este tipo de irracionalidad se produzca realmente. Sin embargo, según algunos autores preferir lo peor no sólo es algo que sucede con frecuencia, sino que no se trata en absoluto de un tipo de irracionalidad. Estos casos sucederían en situaciones como la siguiente:

Imagina que es media tarde; has tenido un buen almuerzo y ahora no tienes hambre; por otro lado, tampoco estas saturado. Disfrutarías con una barra de caramelo o con una Coca-cola si los tuvieras y, de hecho, justo al lado de tu escritorio hay un frigorífico lleno de tales cosas puesto a tu disposición de modo gratuito por la compañía para la cual trabajas. Si eres consciente de todas estas cosas, entonces ¿tomas uno de estos refrigerios y lo consumes necesariamente? Y si no lo haces, ¿se debe esto necesariamente a que temes echar a perder tu cena, a que estas a dieta o a que estas demasiado ocupado? Yo creo que no. Puede ser que, simplemente, tú no sientas la necesidad de consumir estas cosas. Rechazas algo bueno, una satisfacción segura, porque estas perfectamente satisfecho tal y como estas. La mayoría de nosotros nos encontramos a menudo en una situación como esta, y muchos de nosotros haríamos esto a menudo. No somos optimizadores ni maximizadores ilimitados, sino que a veces somos más modestos en nuestros deseos y necesidades. Pero tal modestia, tal moderación, no tienen porque ser irracionales o irrazonables por nuestra parte¹⁴.

En este texto queda absolutamente claro que la alternativa rechaza es la mejor desde todos los puntos de vista. No se trata de no preferir algo que es mejor en algunos aspectos debido a que, desde otros puntos de vista, presenta inconvenientes que hacen que sea en definitiva peor. Simplemente, hemos alcanzado ya un nivel de satisfacción a partir del cual ya no queremos más. Tampoco es que a partir de ese nivel seamos incapaces de discriminar entre mayor y menor satisfacción y, por tanto, seamos indiferentes ante la alternativa de obtener más o menos. Nuestro individuo disfrutaría con el caramelo. Tampoco sucede que el rechazo de un bien que exceda nuestro nivel de satisfacción nos reporte algún tipo de compensación indirecta, haciéndonos sentir, por ejemplo, más a gusto con nosotros mismos. El individuo del que habla Slote admite, en resumen, que tomar una Coca-cola sería lo mejor para él desde cualquier punto de vista, se mire por donde se mire y se le den las vueltas que se le den. No hay ningún truco. Este individuo, simple y llanamente, prefiere lo peor¹⁵.

¹⁴ Slote (85), p.39

¹⁵ Para una discusión detallada de la postura defendida por Slote, ver Apéndice 3.

Podemos resumir este punto del modo siguiente. Hay un modo en el que las preferencias de un agente pueden ser intrínsecamente irracionales. Esta acusación de irracionalidad no puede aplicarse a las preferencias primarias. Sin embargo, si es aplicable a lo que llamaremos "*preferencias secundarias*" o, simplemente "*preferencias*"¹⁰ es decir, a las preferencias entre distintas experiencias deseables e indeseables. Llamemos "*deseable*" a las experiencias placenteras e "*indeseable*" a las experiencias desagradables. Aunque distintos individuos pueden encontrar que distintas cosas son para ellos agradables o desagradables (lo cual sería una cuestión de preferencias primarias), sin que pueda hablarse en estos casos de irracionalidad, si puede hablarse de irracionalidad cuando un individuo elige, sin ninguna razón, la peor de dos experiencias. Supongamos que un individuo A considera que la alternativa X es mejor que Y. Esta preferencia primaria no puede ser irracional aunque todos los demás individuos prefirieran unánimemente Y a X. Pero supongamos que el individuo A establece entre los elementos del conjunto de alternativas {X,Y} una relación de ordenación tal que Y es preferida a X. Entonces, salvo que el individuo pueda presentar alguna razón, su preferencia es irracional.

Las preferencias de un individuo aun pueden ser irracionales en otro sentido distinto, aunque relacionado con el anterior. Para poder ver con claridad este segundo sentido, es preciso analizar por qué un individuo prefiere primariamente determinadas cosas. En general, la deseabilidad de los objetos o de las situaciones está relacionada con alguna característica de estos. Pensemos en un cuadro. Un cuadro puede gustarme por muchas razones: porque es de estilo impresionista, porque es azul o porque al mirarlo siento una determinada sensación. Todas estas razones son distintas respuestas posibles a la pregunta acerca de por qué me gusta el cuadro. Igualmente, una determinada comida puede gustarme, por ejemplo, porque es dulce. La respuesta "porque sí" casi nunca es una respuesta adecuada a este tipo de preguntas. Esto sucede porque al preguntar

¹⁰ Llamamos a las preferencias secundarias simplemente "*preferencias*" por dos motivos. En primer lugar, por que son las preferencias de las que hemos hablado hasta ahora y de las que hablaremos en adelante, por lo que resulta más sencillo llamarlas así y no resulta confuso. En segundo lugar, porque son las preferencias propiamente dichas, en tanto que a partir de ellas el individuo realiza sus elecciones.

por qué nos gusta un cuadro o cualquier otra cosa estamos preguntando por la característica que lo hace deseable para nosotros. Sin embargo, si alguien nos pregunta por qué nos gusta el sabor dulce o el azul, tanto la pregunta como la consiguiente respuesta son de carácter distinto. Si alguien nos hace esa pregunta no nos pregunta por nada acerca del objeto deseado, sino por algo acerca de nosotros mismos, y las respuestas adecuadas a estas preguntas hacen referencia a cosas tales como la costumbre o determinadas características fisiológicas como la necesidad de azúcar. Como este tipo de cosas no suelen ser conscientes y son además el tipo de cosas que habitualmente no requieren justificación racional, consideramos que la respuesta "porque sí" es una buena respuesta. Lo que esto significa es que ante la pregunta de por qué me gusta un cuadro, tanto la respuesta "porque es azul" como la respuesta "porque sí" son siempre respuestas incompletas. Una respuesta completa sería "porque es azul y me gusta el azul", es decir, una respuesta que haga referencia por un lado al aspecto del objeto que lo hace deseable para nosotros y, por otro lado, a nuestra valoración de ese aspecto. Sin embargo, habitualmente consideramos que la respuesta "porque es azul" es suficiente, en primer lugar, porque habitualmente un sabor, un color o una sensación son el tipo de cosas que consideramos que no requieren una explicación posterior en sí mismas y, en segundo lugar, porque, debido a ello, la explicación de "porque eso me gusta" se sobreentiende.

Podemos decir que una cosa nos parece deseable porque la deseamos, pero que siempre la deseamos por algo. La razón por la que algo es deseable no suele ser el propio deseo, pero siempre depende de un deseo. La presencia de una razón para explicar una preferencia es lo que hace posible que las preferencias puedan ser irracionales en un sentido distinto a los considerados hasta ahora. Supongamos que me gusta un cuadro porque es azul. Llamemos X a este cuadro. Supongamos además que el cuadro Y, que también es azul, no me gusta. Entonces esta preferencia es irracional. Naturalmente, alguien podría objetar que no hay nada irracional aquí. Puesto que se trata de preferencias primarias, no es legítimo hablar de racionalidad e irracionalidad. Además, no resulta en absoluto extraordinario que, de dos cuadros azules, uno me guste y otro no. Esta objeción tiene dos partes. La última parte es cierta sin discusión, pero no es relevante. En efecto, no hay nada extraordinario ni irracional en que

un cuadro azul me guste y otro, también azul, no me guste. Pero esto sólo sucede porque normalmente la razón (al menos toda la razón) por la que un cuadro nos gusta no es el color. La cuestión es que si lo fuera, entonces ambos cuadros deberían gustarnos igualmente. Es decir, lo que estamos afirmando en general es que, si X nos parece deseable porque es q , entonces si Y es también q , Y es también deseable. En caso contrario, nuestra preferencia es irracional.

La otra parte de la objeción alude a una afirmación que nosotros hemos mantenido, a saber, que las preferencias primarias no pueden ser irracionales. Sin embargo, nuestra afirmación anterior y la presente no se contradicen. No puede ser irracional que a mí me gusten las duchas frías. Pero si esta ducha me gusta porque es fría, y esta otra también lo es, entonces si me gusta la primera y no la segunda, mis preferencias respecto a las duchas son irracionales. La diferencia es la siguiente. Una preferencia primaria no puede ser irracional. Pero un conjunto de preferencias primarias puede ser irracional sin que ninguno de sus elementos lo sea, simplemente porque la presencia de determinadas preferencias primarias requiere la presencia de otras y excluye la presencia de algunas otras. Una vez entendido esto, podemos decir que en este sentido hay preferencias irracionales al igual que hay preferencias racionalmente exigidas.

Es importante recordar que, aunque estas condiciones se establezcan sobre las preferencias aisladas, hacen necesaria referencia a las relaciones entre preferencias. En efecto, decir en este sentido que una preferencia es irracional no significa que haya nada malo en la preferencia individual, sino en tanto que esta es miembro de un conjunto más amplio de preferencias.

En este sentido, no se identifica la acción racional simplemente como aquella que pone en práctica los mejores medios para un fin dado. Más bien, se exige que estos fines cumplan determinadas condiciones. Estas condiciones se refieren a) al modo en que se forman y mantienen las preferencias, b) a la irracionalidad intrínseca de preferir lo peor y c) a la coherencia de un conjunto de preferencias, e.d., a la exigencia de que la presencia de determinados elementos requiera la de otros y excluya la de algunos otros.

Podemos por consiguiente redefinir la elección racional como la elección derivable de una ordenación sobre el conjunto de las preferencias racionales de un individuo.

Sin embargo, antes de aceptar esta definición, es necesario precisar la extensión de ese conjunto en un sentido que hasta ahora hemos ignorado. En efecto, puede plantearse si el conjunto de preferencias sobre el que se define la elección racional debe contener sólo las preferencias del agente respecto al presente o debe incluir también las preferencias respecto al futuro. Un ejemplo muy simple podría ser el que manejamos al hablar de la necesidad de que las preferencias estuvieran formadas tras una reflexión cuidadosa de los costes de realización de una elección. Decíamos que mi preferencia por ir al cine el sábado por la noche era irracional si se había formado sin considerar que su satisfacción impediría la de otra preferencia, a saber, la de dormir lo suficiente. Ambas preferencias, la de ir al cine y la de dormir, son preferencias que yo tengo ahora, en el momento de hacer la elección, pero son preferencias relacionadas con distintos momentos del tiempo: deseo ir al cine ahora y dormir esta noche.

Nuestras preferencias a veces son tales que no pueden cumplirse a la vez. Por eso hemos introducido la noción de ordenación y de utilidad. En ciertos casos la incompatibilidad de dos preferencias surge de la imposibilidad de satisfacer ambas al mismo tiempo. Un caso de este tipo es mi deseo de ir al cine y al teatro esta noche. No puedo hacer ambas cosas a la vez, de modo que tengo que elegir. Supuesto que estos deseos son racionales, mi elección será racional si consiste en la realización de la alternativa que vaya a rendir más utilidad y, si ambas resultan iguales en este aspecto, realizaré cualquiera de ellas.

Estos casos no plantean ningún problema adicional. Sin embargo, hay otros casos en los que la incompatibilidad de dos preferencias no consiste en la imposibilidad de realizar ambas al mismo tiempo, sino más bien en que la realización de una de ellas ahora anula la posibilidad de que la otra se realice en el futuro. Este futuro puede ser muy cercano, como en el ejemplo anterior. Pero también puede ser bastante lejano. Esto sucede, por ejemplo, con mi deseo de ir al cine todas las noches y mi deseo de no tener arrugas prematuras. Sé que si duermo poco mi piel se estropeará, de modo que tengo que decidir ahora cuál de mis dos deseos

voy a satisfacer. Naturalmente, puede resolverse esta cuestión del mismo modo en que se resuelven los casos de deseos incompatibles en el presente, siempre y cuando se considere que el conjunto de preferencias de un agente se forma teniendo en cuenta todas sus preferencias actuales, sean estas respecto al presente o respecto al futuro¹⁹.

El problema surge cuando se plantea la posibilidad de que las preferencias se jerarquicen por cuestiones de carácter puramente temporal, es decir, según el momento de su realización. Por ejemplo, un agente puede preferir sacarse una muela a sacarse dos, pero puede preferir sacarse dos mañana a sacarse dos hoy, de tal modo que la inversión de la relación de preferencia obedezca únicamente al momento del tiempo en que han de realizarse.

La cuestión es si es racional que consideraciones de carácter temporal afecten a la formación de la jerarquía de preferencias. En el ejemplo anterior, si consideramos que las preferencias deben ser temporalmente neutrales, y supuesto que el agente considere peor sacarse dos muelas que una, entonces sería intrínsecamente irracional preferir sacarse dos muelas mañana a sacarse una muela hoy, pues esto sería preferir lo peor.

Habitualmente se entiende que este fenómeno, por lo demás bastante común, se debe a una falta de información y reflexión adecuadas. Se supone que si el agente reflexiona suficientemente y entiende que un mal o un bien no son menores por ser futuros, supuesto que ocurrirán con certeza y que todas las demás circunstancias permanecerán invariables, dejará de discriminar entre distintos sucesos por la sola razón del momento de su realización. Por ello, es habitual exigir que las preferencias sean en este sentido temporalmente neutrales²⁰.

Con esto hemos acabado el análisis de las condiciones que debe cumplir una elección para ser racional. A partir de este análisis podemos definir la elección racional como la elección derivable de una ordenación sobre el conjunto de las alternativas realizada según las preferencias (racionales) del agente.

¹⁹ Es decir, no solo las preferencias que Harv llama "ahora-para-ahora" sino también las que llama "ahora-para-después" (Harv, 81).

²⁰ Este punto se estudiará con mayor detalle más adelante, concretamente en el capítulo II.

CAPITULO III

SITUACIONES DE ELECCION

La racionalidad práctica ha quedado identificada con la maximización de la utilidad. Este es el modelo de racionalidad generalmente utilizado en la teoría económica clásica, y tiene la ventaja de ser sencillo y bien definido, así como la de gozar de una amplia aceptación.

Sin embargo, este modelo es limitado. Esta limitación surge del hecho de que el modelo de maximización de la utilidad ha surgido del estudio de la elección racional en situaciones de certidumbre, y su aplicación ha estado fundamentalmente restringida a este tipo de situaciones. El problema estriba en que esas situaciones no son las únicas, ni siquiera las más típicas, y el modelo no parece tener éxito cuando se intenta aplicarlo a otro tipo de situaciones.

El objetivo de este capítulo es analizar este problema. Para ello estudiaremos las distintas situaciones de elección en que puede encontrarse un agente y las modificaciones que deben hacerse en el modelo de racionalidad como maximización de la utilidad para que puede aplicarse a tales situaciones.

Es habitual distinguir tres tipos de situaciones de elección, a saber, situaciones de certidumbre, de riesgo, y de incertidumbre. Esta clasificación se establece según el grado de seguridad con que el agente puede esperar que se produzca un determinado acontecimiento como resultado de sus acciones.

Situaciones de certidumbre.

Las situaciones de certidumbre son aquellas en las que el agente puede predecir con seguridad el resultado de las distintas acciones alternativas entre las que puede elegir. Un ejemplo de este tipo de situaciones sería la elección entre dos platos, digamos lasaña verde y pizza cuatro estaciones, en un restaurante. Puedo elegir entre pedir uno u otro de estos dos platos al camarero y sé con seguridad cuál será el resultado en cada uno de los casos, es decir, sé que si pido la pizza el camarero me la traerá y me la comeré, y que comeré lasaña si es este el plato que pido.

En este tipo de situaciones existe una conexión máximamente estrecha entre las acciones del agente y los resultados, de tal modo que puede decirse que la realización de un determinado suceso depende de la acción que ejecuta el agente.

Para comprender adecuadamente las situaciones de certidumbre es conveniente realizar algunas distinciones que no hemos hecho hasta el momento. Estas distinciones son de carácter básico y si no las hemos explicitado anteriormente ha sido porque en el concepto de racionalidad práctica que hemos manejado hasta ahora tales distinciones se desdibujaban.

La primera de estas distinciones se establece entre elegir una acción y elegir un resultado. Hablando con propiedad, el objeto de elección es el conjunto de las distintas acciones alternativas. En un cierto sentido puede hablarse de que elegimos resultados, pero esto sólo quiere decir que preferimos un determinado resultado a otros posibles. Ahora bien, cuando podemos predecir con seguridad el resultado de una acción, elegir una acción es tanto como elegir el resultado de esa acción.

La segunda distinción se refiere al objeto de las preferencias. La relación de preferencia se establece entre un conjunto de resultados alternativos. En un sentido puede decirse que preferimos una acción a otra, pero esto simplemente significa que preferimos el resultado de esa acción

al de la otra¹. Este sentido en el que puede decirse que preferimos una acción adquiere mayor relevancia en las situaciones de certidumbre, pues en ellas la realización de una acción tiene una consecuencia que predecimos con seguridad. Es decir, en estas situaciones la conexión entre una acción y sus consecuencias es de dependencia directa, lo cual hace que la diferencia entre ellas se atenúe.

La tercera distinción es la mencionada anteriormente entre acciones racionales y acciones acertadas o exitosas. Una acción es acertada si consigue realizar el fin propuesto y es racional si es la mejor acción que puede realizar un agente para conseguir un fin dado. Sin embargo, en situaciones de certidumbre ambas cosas coinciden siempre. De nuevo, esto es debido a la conexión que existe en estas situaciones entre la realización de una acción y la consecución de un fin determinado que se sigue con seguridad de la acción. La coincidencia en una misma acción del carácter racional y del éxito que se da en situaciones de certidumbre es tan grande que de hecho podemos juzgar la racionalidad de una acción a partir de su resultado y viceversa.

Esto significa que en las situaciones de certidumbre la identificación de la acción racional con aquella que maximiza la utilidad es perfectamente aplicable. Y significa también que ~~per~~ la determinación de cuál es la acción racional en una situación dada depende únicamente de las preferencias del agente.

Sin embargo, no todas las situaciones de elección son de certidumbre. Y, en esas otras situaciones, la aplicación directa del modelo de racionalidad que hemos utilizado hasta ahora no resulta tan satisfactorio.

¹ En este contexto, hablamos de "acciones" y "resultados de acciones" en el sentido de medios y fines respectivamente. Empleando la terminología de medios y fines es más claro que lo que en una ocasión es un medio puede ser en otra un fin y viceversa.

Supongamos que yo prefiero tomar una clase de baile a ir al cine. Tomar clases de baile e ir al cine son aquí fines, para cuya realización yo debo adoptar determinados medios, como por ejemplo pagar una considerable cantidad de dinero y emplear bastante tiempo en el caso de bailar. Sólo en sentido impropio puede decirse que yo prefiero esos medios. Al mismo tiempo, bailar puede ser considerado como un medio para un fin, digamos la salud y la belleza física. En este sentido, la preferencia por bailar sería la preferencia por un medio y, por tanto, sólo sería una preferencia derivada.

La cuestión no está en que algo sea un medio o un fin. Esta discusión es irrelevante y, además, puede resultar interminable. Lo relevante es que en este sentido de medios y fines las preferencias se establecen con relación a los resultados de las acciones, *i.e.*, los fines.

Situaciones de riesgo e incertidumbre

En las situaciones de riesgo e incertidumbre el agente no puede predecir con seguridad cuál será el resultado de sus acciones. Contrariamente a lo que ocurre en las situaciones de certidumbre, donde cada acción alternativa que el agente puede elegir está asociada con un único resultado, en los casos de riesgo e incertidumbre cada una de las acciones alternativas está asociada con un conjunto de posibles resultados, de tal modo que, si bien el agente sabe que uno de ellos se realizará, no sabe con certeza cuál será éste.

Las situaciones de riesgo se distinguen de las de incertidumbre en que en las primeras el agente conoce las probabilidades objetivas², aso-

² El concepto de probabilidad es utilizado habitualmente en tres contextos diferentes, a los que subyacen acepciones distintas del concepto mismo.

Tradicionalmente (Laplace), la probabilidad ha sido asociada al tratamiento de los sucesos aleatorios, típicamente juegos de azar. En una situación aleatoria, la única información relevante para predecir la ocurrencia de uno u otro suceso (jugada) son las reglas del juego. Con ellas, cada jugador puede establecer el universo de sucesos posibles y, apoyándose en que cada jugada tiene igual oportunidad de acontecer que el resto, asignar una probabilidad a cada suceso. El cálculo efectivo se basa en la llamada 'regla de Laplace': situaciones favorables / situaciones posibles. Para distinguirla de otras, hablamos en este caso de probabilidad a priori, por ser su cálculo una deducción a partir de la descripción de las condiciones de la situación aleatoria. Las asignaciones extremas '1' y '0' corresponden, respectivamente, al suceso universal y al suceso imposible.

Además de las situaciones aleatorias, la probabilidad se aplica también a situaciones en las que, por muy diversas razones, no podemos disponer de toda la información relevante al caso. Se recurre entonces a nuestros estadísticos sobre la población a estudiar, o a la repetición sucesiva, en idénticas condiciones, del experimento que ha producido el fenómeno observado. La estabilidad de la distribución de los sucesos según se amplía la muestra o se acumulan los experimentos conduce a la siguiente definición de probabilidad: la probabilidad de un suceso es la frecuencia límite a la que tienden las frecuencias observadas según ampliamos el número de casos. En un extremo, afirmar la probabilidad '1' equivale a sostener la determinación causal estricta de las condiciones iniciales sobre el suceso observado. En el otro, el grado '0' de probabilidad supone la exclusión causal (igualmente estricta) de los antecedentes sobre el suceso en cuestión. El valor '1/2' representa la irrelevancia causal de las condiciones iniciales sobre el fenómeno. A esta noción de probabilidad como frecuencia a largo plazo se la denomina también probabilidad a posteriori.

Por último, hablamos de probabilidades para referirnos a nuestro grado de conocimiento sobre la ocurrencia de un suceso, a la apreciación personal de su aparición. Naturalmente, esta apreciación probabilística se fundamenta en ciertos conocimientos (más o menos exactos) sobre los que hacer nuestro juicio. Sin embargo, los valores no son una traducción directa de esta información objetiva (reglas de juego, estadísticas o frecuencias observadas), sino una ponderación personal de aquellas. En esta 'tercera acepción', la probabilidad se identifica con el grado de certeza o nivel de confianza subjetiva en el acontecimiento de determinado suceso. Se trata por tanto de probabilidad subjetiva, en la que los valores extremos '1' y '0' representan, respectivamente, certeza absoluta e ignorancia completa.

Teniendo en cuenta estos conceptos de probabilidad y la definición de las situaciones de riesgo e incertidumbre, diremos que las primeras se dan cuando el sujeto tiene un conocimiento completo de las

ciadas con cada uno de los posibles resultados de una acción, mientras que en las situaciones de incertidumbre estas probabilidades son desconocidas³.

reglas del juego en los casos de probabilidad aleatoria o bien cuando tiene un conocimiento estadístico tan completo como sea posible en los casos de probabilidad a posteriori (para una situación de riesgo en el que se haya involucrado el concepto de probabilidad estadística, ver el ejemplo del huevo podrido y el granjero científico en Luce y Raiffa 57, p.277). Las situaciones de incertidumbre se darán cuando el agente carece completa o parcialmente de este conocimiento. Naturalmente, esto sucederá con mayor frecuencia en los casos de probabilidad estadística y sólo cuando las reglas del juego son extremadamente complicadas en el caso de la probabilidad aleatoria. Por lo general, aquí se utilizarán casos del primer tipo de probabilidad para ilustrar situaciones de riesgo y del segundo para casos de incertidumbre. Es decir, asumiremos por lo general que es el desconocimiento de la probabilidad estadística lo que da lugar a la aplicación de probabilidades subjetivas.

³ Este tipo de situaciones no es en absoluto inusual, más bien al contrario. Podemos ilustrar estas situaciones con una variante del ejemplo que utilizamos en el capítulo anterior. Supongamos que yo puedo elegir entre ir al cine a ver "El último emperador", ir al teatro a ver "Enrique IV" y quedarse en casa estudiando. Llamaremos a estas acciones alternativas A, B y C respectivamente, y a los resultados de ver la película, ver la función y estudiar, a, b y c. Supongamos también que la alternativa se me plantea a las 7 de la tarde del sábado, y que "El último emperador" acaba de ser estrenada con gran éxito de público mientras que "Enrique IV" lleva varios meses en cartel.

Siendo esta la situación en la que debo hacerse la elección, nos encontramos con un caso en el que no podemos predecir con seguridad el resultado de nuestra elección, o.d., no es un caso de certidumbre. Más bien al contrario, cada una de las acciones que yo puedo realizar está asociada con varias consecuencias posibles. Así, la acción de ir al cine puede tener como resultado que vea la película o que no la vea, y la acción de ir al teatro puede tener como consecuencia que presencie la función o que no lo consiga. Naturalmente, este ejemplo es una simplificación de una situación real en la cual, aparte de los posibles resultados señalados de cada acción alternativa, existen otros muchos igualmente posibles. Si me quedo en casa a estudiar se me puede caer el techo encima, o me pueden llamar de un concurso televisivo que me ofrezca un premio. Si salgo de casa para ir al cine o al teatro puedo encontrar al hombre de mis sueños, o puede que me atraquen. Lo que esto significa es que en la vida real cualquier cosa que hagamos puede tener multitud de consecuencias posibles, algunas de las cuales no podemos siquiera prever. Más adelante en este capítulo veremos que implicaciones tiene este hecho para el propio concepto de maximización. Por el momento, y a fin de presentar el problema del modo más simple posible, nos limitamos a asociar con cada acción alternativa aquellas consecuencias que no sólo son posibles sino probables, entendiendo ahora este término en el sentido coloquial (implicado en frases como "es posible que el se hunda el suelo bajo mis pies, pero no es probable").

En este caso si bien el conjunto de alternativas entre las que tenemos que elegir sigue siendo el conjunto (A,B,C), el conjunto de posibles resultados se amplía en un miembro, a saber, el resultado negativo de las acciones A y B que consistiría en darse un paseo en balde. Llamemos d a este resultado. Supongamos que este conjunto de posibles resultados está ordenado por la relación R, de modo que quedan jerarquizados en el orden a,b,c,d.

La situación de elección planteada en este ejemplo es de incertidumbre, pues yo no conozco la probabilidad objetiva asociada con los posibles resultados de cada una de las acciones.

En una situación de riesgo si se conoce este dato. Veamos un ejemplo. Supongamos que se presenta a un concurso de televisión y como recompensa a la buena actuación se me ofrece la posibilidad de conseguir un premio. Se colocan ante mí dos urnas opacas, de modo que no puedo ver lo que hay en el interior. Dentro de cada una de las urnas hay ocho bolas, blancas y rojas, en la siguiente proporción: en la urna A hay seis bolas rojas y dos blancas y en la urna B cinco rojas y tres blancas. Yo puedo sacar una bola de una de las urnas. Si se decide por la urna A los premios consistirán en una semana en

El concepto de acción racional que se utiliza en situaciones de certidumbre no puede ser aplicado con éxito en estos casos. En las situaciones de certidumbre vemos que la acción racional estaba determinada únicamente por las preferencias del agente hacia los resultados de cada una de las acciones alternativas. En los casos que no son de certidumbre no puede suceder así. Supongamos un conjunto $\{a, b, c, d\}$ de alternativas jerarquizadas con el orden a, b, c, d . Si la situación de elección fuera de certidumbre tendríamos un conjunto de acciones alternativas A, B, C, D cada una de ellas asociada con uno de los resultados jerarquizados según la relación de preferencia. En esta situación, la acción racional sería A , ya que sería esta la acción que maximizaría la utilidad. Ahora bien, en casos de riesgo e incertidumbre cada acción alternativa está asociada con más de un resultado posible. Por ejemplo, de la acción A puede seguirse el resultado a o el resultado d , y de B pueden seguirse b o d . No hay ninguna acción alternativa de la que se siga con seguridad el resultado que rendiría una mayor utilidad al agente, e.d., de la que se siga a . Todo lo que tenemos es una acción que *puede* tener como resultado a , pero que también *puede* tener como resultado d .

Para elegir racionalmente en estos casos no sólo deben tenerse en cuenta las preferencias del agente respecto a los resultados alternativos, sino también las probabilidades de que estos resultados se sigan de una determinada acción. Es habitual en la literatura especializada el uso de una metáfora que recoge la exigencia de que se consideren también las probabilidades. Esta metáfora consiste básicamente en identificar las distintas acciones alternativas con apuestas o loterías y sus posibles resultados con los posibles premios. Esta metáfora es perfectamente aplicable y resulta sumamente útil. De hecho, una apuesta en sentido habitual es una elección arriesgada o incierta según los casos.

Tenemos por tanto que la elección racional debe hacerse en estos casos en función de las preferencias del agente y de las probabilidades.

Sevilla si saca una bola roja y si elijo jugar con la urna B ser de una semana en Venecia, siempre y cuando sea roja la bola. En las dos urnas, si saca una bola blanca no conseguirá nada.

En esta situación yo conozco las probabilidades objetivas de cada uno de los dos posibles resultados de cada acción alternativa: si saca una bola de la urna A se que puedo conseguir un viaje a Sevilla con una probabilidad de 0,75, y si saca una bola de la urna B la probabilidad de ganar el viaje a Venecia es de 0,65.

La idea intuitiva no representa ningún problema, pues esta exigencia recoge nuestras ideas preteóricas sobre la acción racional. En efecto, la decisión racional en el caso del sábado por la noche (ver nota 3) puede variar dependiendo de si es muy probable o poco probable que haya entradas en el cine y en el teatro.

Sin embargo, la cuestión es bastante más complicada de lo que en principio pueda parecer y la necesidad de tener en cuenta las probabilidades origina algunos problemas que no aparecían anteriormente. Estos problemas salen a la luz en cuanto se considera uno de los casos con mayor detenimiento.

Supongamos que las probabilidades de encontrar entradas para el cine son escasas mientras que es bastante probable que haya entradas para el teatro. Es decir, si voy al cine es muy probable que el resultado sea d (darme un paseo en balde) y muy poco probable que consiga a (ver "El último emperador"). Si por el contrario voy al teatro es poco probable que el resultado sea d y bastante probable que sea b. Podría pensarse entonces que, en principio, la acción racional es B. Pero esto no es así necesariamente. Que lo sea o no depende de la intensidad con la que prefiera a sobre b. Y depende también del grado de probabilidad de cada uno de estos sucesos. Si yo prefiero muy débilmente ver la película a ver la función pero me parece horrible darme un paseo sin sentido y la probabilidad de que esto último suceda es muy elevada si voy al cine y muy escasa si voy al teatro, desde luego la acción racional es ir al teatro. Pero si, por el contrario, mis deseos de ver la película son muy grandes y soy casi indiferente entre las otras alternativas, todas las cuales me parecen poco apetecibles, entonces lo más racional será seguramente ir al cine, ya que así tengo al menos la posibilidad de pasar una noche agradable.

Por lo tanto, en las situaciones de riesgo e incertidumbre necesito contar no sólo con una ordenación de los resultados alternativos sino con una medida mucho más precisa, con una medida del intervalo entre cada una de las alternativas y las demás*.

* Podría pensarse que es suficiente con saber si una determinada alternativa es mucho más preferida que otra, o bastante o poco. Esto no es así dado que la intensidad de las preferencias debe considerarse junto a otro factor, la probabilidad de su ocurrencia. En primer lugar, no parece posible encontrar un modo de decidir sobre esta base. Supongamos que *m* se desea con mucha intensidad, *h* con poca

Por todo ello se necesita una medida cardinal sobre las preferencias. Una medida cardinal nos dirá no sólo que alternativa tiene más utilidad sino también *cúanta* utilidad tiene cada alternativa. Una vez que contemos con esta medida será posible definir la conducta racional en situaciones de riesgo e incertidumbre.

En situaciones de certidumbre las preferencias se establecen únicamente respecto a los resultados. La segunda distinción que hicimos anteriormente (p.34) respecto al objeto de las preferencias señalaba este hecho, explicando cómo en casos de certidumbre la preferencia por una acción (o dicho de un modo menos ambiguo, la elección de una acción) derivaba directamente de las preferencias sobre sus resultados y era idéntica a ella. Sin embargo, esto no ocurre en situaciones de riesgo e incertidumbre. En estas situaciones el objeto originario de preferencia siguen siendo los resultados pero, puesto que no hay acciones que conduzcan a estos con seguridad, las preferencias por los resultados no se traducen directamente en elecciones de acciones. No obstante, en estos casos, para mantener la conexión necesaria entre elección y preferencias, las decisiones deben seguir dependiendo de alguna manera de las preferencias por los resultados. Por tanto, en situaciones de riesgo e incertidumbre la elección de la acción a realizar estará en función de la probabilidad con que estas acciones conducen a los resultados, de tal modo que elegir una acción A es preferir la probabilidad p de que el resultado a se produzca². Es decir, en las situaciones de riesgo e incertidumbre, al contrario de lo que sucedía en las situaciones de certidumbre, sí tiene sentido hablar de preferencia por las acciones como distinta a la preferencia por los posibles resultados. De hecho, en estas situaciones esta preferencia establecida sobre las acciones (e.d., el hecho de preferir

y c con poquísima y que con la acción A la probabilidad de conseguir a es escasa y la de d grande, mientras que con la acción B la probabilidad de b es moderadamente grande y la de d no mucho. No parece factible encontrar una función que ponga en relación "mucho" con "escasa" o "escasa" con "poquita". Además, y en segundo lugar, en ocasiones será necesaria una diferencia en intensidad más sutil que la que va de "mucho" a "bastante", así como también lo será contar con una medida de probabilidad precisa. Esto sucede porque, al entrar en consideración dos factores, es posible que una diferencia relativamente pequeña en una de ellas pueda alterar la decisión.

² Por ejemplo, en el caso del concurso (nota 3), preferir la acción B a la acción A es preferir la probabilidad 0.625 de conseguir un viaje a Venecia a la probabilidad 0.75 de conseguir un viaje a Sevilla.

realizar una acción a realizar otras), es la determinante y, por tanto, la utilidad determinante será la medida de estas preferencias. Esta medida recibe el nombre de *utilidad esperada* y se obtiene sumando los productos que resultan de multiplicar la utilidad de cada resultado posible de una acción por la probabilidad de su ocurrencia.

Por tanto, en situaciones de riesgo e incertidumbre cada acción alternativa tiene asociada una utilidad. En el capítulo II analizamos una serie de condiciones que debía cumplir el conjunto de preferencias de un agente para que la elección racional fuera posible. Una vez que se cumplían tales condiciones era posible identificar la elección racional como aquella que maximizaba la utilidad. Hemos visto que este concepto de acción racional no es aplicable en casos de riesgo e incertidumbre y que es necesario definir de otra forma la acción racional en estas condiciones. Un primer paso en este sentido consiste en desplazar la preferencia desde los resultados a las acciones y definir una medida sobre estas preferencias, la utilidad esperada. Ahora bien, para que esta medida pueda ser utilizada deben imponerse algunas condiciones al conjunto de preferencias sobre las acciones del mismo modo que antes las imponíamos al conjunto de preferencias sobre los resultados.

Las condiciones anteriores siguen siendo necesarias, puesto que el concepto de utilidad entra en la definición de la utilidad esperada. Además, estas condiciones deben cumplirse para que el conjunto de preferencias pueda ser ordenado jerárquicamente según una relación de ordenación, lo cual, como vimos, es un requisito necesario para que pueda haber elección racional. Aparte de estas condiciones, el conjunto de preferencias sobre las acciones debe cumplir otras dos:

1. *Sure-thing principle*⁴. Supongamos que se nos presentan dos boletos de lotería, L y L' tales que L tiene como posibles premios a_1, a_2, \dots, a_n , con las probabilidades respectivas p_1, p_2, \dots, p_n , y L' es exactamente igual salvo que el premio a_1 es sustituido por otro premio a'_1 y

⁴ En algunas presentaciones de las condiciones de racionalidad en situaciones de no certidumbre se sustituye este principio por la condición de monotonicidad. Esta es una versión más fuerte de este principio, pues exige no sólo que el boleto L' tenga al menos tanto valor como el boleto L , sino que tenga más (ver, por ejemplo, Gauthier 86, p.44 y Marsanyi 77c, p.33).



que preferimos a_1 a a_2 . Entonces, el valor del boleto L' será al menos igual al del boleto L'' .

Veamos un ejemplo. Supongamos que en el concurso de televisión que mencionábamos anteriormente nos dan a elegir de nuevo entre dos urnas con ocho bolas cada una. En la urna A hay cinco bolas rojas, dos blancas y una azul, de entre las que tenemos que escoger una al azar. Si sacamos una bola roja el premio será un viaje a París, a Mallorca si sacamos una bola blanca y a Sevilla si sacamos una bola azul. La urna B es exactamente igual, sólo que en esta si sacamos una bola roja el premio es un viaje a Venecia. Supuesto que yo prefiera Venecia a París, el principio nos dice que debemos preferir jugar con la urna B a jugar con la urna A.

2. *Continuidad.* Supongamos que hay tres resultados: a, b, c , tales que el agente prefiere a a b y b a c . Entonces existe una probabilidad p tal que el agente será indiferente entre conseguir b con seguridad y un boleto de lotería que tenga como posibles premios a y c con las probabilidades p y $1-p$ respectivamente. Es decir, hay un valor p tal que $p * U_a + (1-p) * U_c = U_b$.

Supongamos, por ejemplo, que hay tres resultados, a, b, c , tales que $U_a = 100$, $U_b = 10$ y $U_c = 1$. A partir de estos valores podemos determinar un valor de p tal que seamos indiferentes entre conseguir b con seguridad y un boleto de lotería que nos de la oportunidad p de conseguir a y la oportunidad $1-p$ de conseguir c . Es decir, podemos hallar un valor de p tal que $p * 100 + (1-p) * 1 = 10$. En este caso el valor de p es 0.0909 ². Es decir, seríamos indiferentes entre esta lotería y la seguridad de conseguir b .

² Esta es la formulación del principio para situaciones de riesgo. Para situaciones de incertidumbre la formulación es ligeramente distinta. La diferencia se debe a que en condiciones de riesgo se opera con probabilidades positivas, mientras que en casos de incertidumbre cada premio posible se asocia con una probabilidad subjetiva que puede ser igual a cero. En este caso, si el premio que distingue ambas loterías tiene como probabilidad subjetiva cero, ambas loterías serán indiferentes. Para una formulación precisa ver Harsanyi (1976) p.383.

³ Esto no es exacto. El valor de la lotería es 9.9991. La diferencia se debe a que no hemos calculado p con exactitud, pues sólo hemos operado con los cuatro primeros decimales para simplificar. Si seguimos operando con más decimales el resultado se irá aproximando a 10. Por ejemplo, si asignamos a p el valor 0.090909, el valor del boleto de lotería ascendería a 9.999991, etc.

La idea intuitiva subyacente a esta condición queda bien expresada en el concepto de "punto de inversión". Dado que a es preferida a b y esta a c , es de esperar que una lotería que tenga como premios a a con la probabilidad p y a c con la probabilidad $1-p$ sea preferida a la seguridad de obtener b si p es cercano a uno, y que, al contrario, se prefiera la seguridad de b si el valor de p es cercano a 0. Siendo esto así, resulta intuitivo suponer que habrá un valor intermedio de p en el que la preferencia entre la lotería y la seguridad de b se invierte. Este valor de p será el punto de inversión en el cual el agente será indiferente entre ambas alternativas¹⁰.

Estas dos condiciones, junto con las establecidas en el capítulo II, nos permiten determinar la utilidad de las acciones en situaciones de riesgo e incertidumbre. En efecto, si las preferencias de un individuo satisfacen estas condiciones de racionalidad entonces es posible aplicar una medida cardinal a sus preferencias sobre los distintos resultados alternativos, y definir posteriormente la utilidad de cada acción alternativa como su utilidad esperada¹¹.

Una vez que tal medida ha quedado definida, puede decirse que la acción racional se identifica con la acción que maximiza la utilidad esperada. Esta definición se establece en el estudio de las situaciones de riesgo e incertidumbre pero puede hacerse extensiva con facilidad a los casos de certidumbre. Estos pueden interpretarse como casos límite en los que la probabilidad con la que una acción conduce a un resultado determinado es igual a 1, con lo cual la utilidad esperada de la acción es igual a la utilidad inicial de su resultado.

Este es el concepto de racionalidad práctica defendido por la teoría Bayesiana de la decisión. La defensa de este concepto se sigue de la aceptación de las condiciones de racionalidad exigidas por esta teoría. Como señala Harsanyi (Harsanyi 77b, p.382) la teoría Bayesiana se sustenta o fracasa con la validez o invalidez de estas condiciones. Nosotros las

¹⁰ Luce y Raiffa 37, p.27.

¹⁰ Algunos autores han objetado a esta condición que en algunos casos conduce a resultados absurdos. Por ejemplo, Gauthier (86, p.45) utiliza un caso donde tales resultados absurdos parecen seguirse. Para un análisis y discusión de esta crítica, ver Apéndice 4.

¹¹ Para una discusión detallada del método de cardinalización, ver Apéndice 5.

hemos analizado y encontrado aceptables, lo cual nos compromete con la aceptación de este concepto de racionalidad práctica. Sin embargo, algunos autores sostienen que el modelo de racionalidad como maximización se sustenta sobre unas exigencias tales que no puede aplicarse en las situaciones reales medianamente complicadas. No se trata por tanto de una objeción a la teoría sino más bien a su aplicabilidad. Por ello, han propuesto un modelo alternativo, el modelo de satisfacción, que pretende responder de una manera más adecuada al conocimiento y las capacidades de los agentes reales. A la discusión de este modelo dedicaremos el resto de este capítulo.

Satisfacción y maximización

Resumiendo todo lo dicho hasta este momento, podemos afirmar que la conducta racional de un individuo depende de

- a) las acciones alternativas que se le presentan
- b) los resultados de esas acciones y
- c) sus preferencias por esos resultados.

En los casos reales medianamente complejos, la determinación de estos factores plantean un problema al concepto de racionalidad como maximización, problema que surge de las demandas que este concepto impone sobre el agente. Concretamente, el agente debe ser capaz de

1. determinar las retribuciones que obtendrá de cada una de las posibles consecuencias
2. realizar una ordenación de tales retribuciones
3. conocer con certeza cuales serán los resultados de las distintas acciones alternativas o bien asignarles una probabilidad determinada.

Algunos autores han expresado sus dudas sobre la posibilidad de que estas exigencias puedan ser satisfechas por los agentes reales en situaciones complejas y han propuesto como solución una teoría sobre la elección racional que pueda ajustarse más a las capacidades reales de los individuos. Esta teoría alternativa propone un concepto de racionalidad como satisfacción y no como maximización¹². Estos conceptos difieren en que "mientras que un agente maximizador sólo obtiene éxito si alcanza la solución correcta, un agente satisfactor se contenta con una suficientemente buena"¹³. Con este fin, Simon propone una serie de simplificaciones de las mencionadas exigencias¹⁴. Para ver con más exactitud el alcance de la noción de satisfacción, así como su relación con el concepto de maximización, y el carácter de estas simplificaciones, veamos en qué situaciones se aplica y qué problemas exactamente se propone solucionar.

Imaginemos una pareja que esta buscando un piso. Si pudieran sentarse delante de una lista "manejable" de opciones, podrían realizar una

¹² No confundir este concepto con el mantenido por Slovic al cual nos referimos en el capítulo II.

¹³ Hollis 87, p.114.

¹⁴ Simon 69

elección maximizadora. Pero habitualmente no disponen de una lista así. Puede pensarse que si la pareja contara con una información mayor, e.d., si dispusiera de una lista completa, sería capaz de obtener un mayor rendimiento. Sin embargo, esto no es así. En primer lugar, una lista ideal que contuviera realmente todas las opciones no sería "manejable" ya que contendría un número de casas mayor que el que pueden esperar visitar (supongamos, muy realísticamente, que hay que visitar una casa para hacerse una idea medianamente adecuada de cómo es).

Este primer problema puede ser analizado en términos de costes y beneficios. Supongamos que para visitar todas las casas de la lista ideal, la pareja tendría que dejar de trabajar y dedicar todo su tiempo a esta tarea. Obviamente, el costo de adquisición de esta información sería mucho mayor que sus posibles beneficios. De modo que, en términos de maximización de la utilidad, nuestra pareja hará bien en hacer una selección previa que reduzca su lista a unas dimensiones razonables, aún a riesgo de dejar fuera de la lista algo mejor que lo que se queda dentro.

En segundo lugar, cada día salen al mercado nuevas casas, de modo que esa lista ideal ni siquiera podría llegar a redactarse. No se trata ya, como en el caso anterior, de un problema del costo de la información, sino de la imposibilidad de tener una información absolutamente completa. Naturalmente, la pareja quiere comprar un piso ahora, de modo que para ellos no resultan relevantes las casas que puedan salir al mercado dentro de un año o más. Pero "ahora" no es necesariamente "hoy", sino que puede ser mañana, la semana que viene o dentro de un mes. Parece entonces que harían bien en esperar, pongamos, un mes más. Sin embargo, esto no es así, pues, se pongan el plazo que se pongan siempre es posible que la casa de sus sueños salga al mercado mañana. De modo que la probabilidad de que encuentren algo mejor una vez que ya haya comprado una casa es igual si toman la decisión hoy o dentro de un mes¹². Por tanto, la pareja hará muy bien en tomar su decisión ahora, siempre y cuando encuentren en el mercado algo que les parezca suficientemente bueno.

¹² Naturalmente, salvo que tengan razones para suponer que en un plazo determinado de tiempo habrá mejores opciones, por ejemplo, si saben que dentro de tres meses acabaran de construir unos pisos maravillosos en el barrio que más les gusta.

Por último, nuestra pareja puede tener dificultades para ordenar su lista, incluso si han conseguido una lista manejable. En efecto, para tomar su decisión, tienen que valorar factores muy distintos, tales como cercanía al centro o al lugar de trabajo, amplitud de la casa y distribución del espacio, luminosidad etc. Es fácil ordenar las alternativas con respecto a uno de estos factores, pero la cuestión puede no resultar tan sencilla cuando se trata de decidir si es mejor 10 metros cuadrados más o nueve estaciones de metro menos.

Las simplificaciones propuestas por Simon tienen como objeto reducir las exigencias del modelo de racionalidad como maximización. En concreto, Simon propone

1. frente a la exigencia de determinar con exactitud las retribuciones que se obtendrán de cada una de las alternativas, exigir solamente que el agente sea capaz de determinar la retribución de cada uno de los resultados posibles (la compra de una u otra casa) con dos valores (1,0), que pueden ser interpretados como "satisfactoria" e "insatisfactoria"¹⁴.

2. reducir la exigencia de una ordenación completa de las retribuciones de cada alternativa. Para ello propone sustituir la función de utilidad U , que asigna un determinado valor a cada una de las alternativas, por una función vectorial V , donde cada alternativa esta asociada a un vector (V_1, V_2, \dots) . Esta función vectorial se aplicaría:

- a) en los casos en los que el agente tiene que decidir entre varias alternativas utilizando más de un criterio, e.d., valorando más de un factor. Por ejemplo, nuestra pareja asignaría a cada casa de su lista reducida un valor vectorial (V_1, V_2, V_3) , correspondiente a la "puntuación" de la casa respecto "cercanía del centro", "amplitud" y "luminosidad" respectivamente, sin verse obligados a utilizar una medida común para todos esos factores. Esta función también se aplicaría

- b) en los casos de no certidumbre, en los cuales cada conducta alternativa esta asociada a un conjunto de resultados posibles. En estas situaciones, la función vectorial nos permite asociar cada alternativa

¹⁴ Alternativamente, en otras situaciones pueden manejarse tres valores (1,0,-1) que pueden interpretarse como "victoria", "empate" y "perdida" (Simon 69, p.105.)

con una única consecuencia con un valor vectorial (V_1, V_2, \dots, V_n) cuyos componentes representan los valores de cada resultado posible¹⁷.

Con estas simplificaciones sería posible

1- determinar el punto en que debe cesar la búsqueda de más información. Este punto se alcanzará en el momento en que se disponga de un resultado satisfactorio,

2- decidir en los casos en los que el agente no cuenta con una ordenación completa de los resultados o en los casos en los que estos son inciertos mediante la misma regla de decisión. Tal regla diría

-busca un subconjunto S' del conjunto de resultados S tal que $V(s)$ sea satisfactorio para todo s en S' y entonces

- busca un elemento a del conjunto de acciones alternativas A tal que el resultado s de a esté en S' .

Volviendo a nuestro ejemplo, la pareja realizará su elección buscando una casa que sea satisfactoria en todos los factores relevantes y dará por concluida la busca en el momento en el que se encuentren con una casa así.

Indudablemente, nuestra pareja no ha comprado la "mejor casa posible", sino una suficientemente buena. Ha realizado una elección satisfactoria. Sin embargo, no puede decirse que no hayan maximizado con su elección. No han visitado todas las casas en venta, sino sólo las que han pasado una selección previa. Probablemente, en su lista están las mejores. Aun así, es posible que no hayan considerado alguna que es mejor que las que están en la lista restringida. Pero esta eventualidad resulta bastante improbable, al menos si han realizado la selección con buenos criterios, lo que podemos dar por supuesto. En cualquier caso, el ahorro en tiempo y esfuerzo compensa con creces ese riesgo. Han comprado la casa hoy y, por tanto, tienen que contar con una determinada probabilidad de que mañana salga al mercado otra mejor. Pero ese riesgo es inevitable, pues la misma probabilidad existe en cualquier momento. Habría que comprar una casa tan buena que no pueda haber otra mejor para evitar ese riesgo. Pero la "mejor casa" no existe. Lo único que pueden hacer es com-

¹⁷ Otra posible aplicación de las funciones vectoriales son los casos en los que la decisión ha de ser tomada, no por un agente individual, sino por un grupo de agentes cuyas preferencias no tienen por que coincidir (Simon 69, p.108).

prar la mejor casa posible. Y puesto que tienen que comprar una casa en algún momento, esto significa que tienen que comprar la mejor posible en ese momento. Nuestra pareja no ha comprado la mejor casa. Pero dadas las circunstancias, conformarse con una casa suficientemente buena es la mejor manera de maximizar.

Este análisis alternativo de la decisión de la pareja en términos de maximización no puede utilizarse a menos que contemos con una medida única que permita ordenar los distintos resultados y evaluar los costes de adquisición de información. La disponibilidad o no de una medida de este tipo parece ser lo que en último término determina la utilización de uno u otro modelo de racionalidad

"Si las retribuciones fueran medibles en dinero o en términos de utilidad, y si los costes del descubrimiento de nuevas alternativas fuera también medible de modo similar, podríamos reemplazar el orden parcial de las alternativas (...) por una ordenación completa (una ordenación en términos de la suma ponderada de las retribuciones y los costes del descubrimiento de alternativas). Entonces podríamos hablar del grado óptimo de persistencia de un organismo [e.d., del grado óptimo en que el agente debe persistir en su búsqueda de mejores alternativas]- podríamos decir que el organismo más persistente es más racional que el otro o viceversa. Pero el argumento central de este artículo es que el agente en general no conoce tales costes, ni tiene un conjunto de pesas con las que comparar los distintos componentes de una retribución múltiple. Es precisamente a causa de estas limitaciones en sus conocimientos y capacidades que estos modelos de racionalidad menos globales descritos aquí resultan importantes y útiles"¹⁸

De hecho, los defensores del modelo de maximización sostienen que las exigencias aludidas (ver p.45) pueden ser satisfechas puesto que se dispone de una función de utilidad cardinal¹⁹. En cualquier caso, conviene tener en cuenta que los conceptos de "satisfacción" y "maximización" no son "completamente diferentes ni, mucho menos, antitéticos"²⁰.

¹⁸ Simon 69, p.112.

¹⁹ Para un análisis del método de cardinalización, ver Apéndice 5. Para una deducción completa de la función de utilidad cardinal, ver Harsanyi 77c, pp.32-41.

²⁰ Simon 69, p.112.

CAPITULO IV

RACIONALIDAD ESTRATEGICA

En los capítulos anteriores hemos analizado el concepto de acción racional en situaciones en las cuales el resultado depende de la acción de un sólo individuo. Sin embargo, no todas las situaciones de elección son de ese tipo. Por el contrario, en muchas situaciones el que un agente consiga o no determinado resultado depende no sólo de sus propias acciones sino también de las acciones de otros.

En un sentido, la satisfacción de las preferencias de un agente casi siempre depende de las acciones de los otros. Por ejemplo, si yo quiero ir al cine, la satisfacción de este deseo depende no sólo de que yo realice ciertas acciones, como vestirme, coger un autobús, sacar una entrada, etc. También depende de que otra persona conduzca el autobús, de que otra venda entradas, etc. Pero a pesar de esto, hay un sentido importante en el que puede decirse que la consecución del resultado deseado depende sólo de mí, a saber, en el sentido de que el conductor del autobús y la taquillera van a cumplir su cometido con total independencia de mi decisión y de mis acciones. Es decir, que yo consiga ver la película depende del conductor y de la taquillera como depende de que haya cines y de que se haya filmado la película, o de que yo tenga gafas para corregir la miopía y de que se cumplan determinadas leyes naturales. Todo esto y muchas otras cosas son condiciones de posibilidad de que yo vea la película. Pero todas ellas se cumplen con independencia de mis acciones y de una manera invariable. Mis acciones son la única variable que resulta decisiva para la consecución de ese resultado.

En otras muchas situaciones mi conducta no es la única variable que determina la consecución de un resultado. En esas situaciones el resultado depende de la actuación de más de un agente, cada uno de los cuales se comporta de acuerdo con sus intereses, intereses que no sólo no coinciden siempre con los míos, sino que en ocasiones son incluso incompatibles. Y, lo que es más importante, esos otros agentes deciden su conducta teniendo en cuenta la mía. En este sentido, mis acciones no se realizan en un entorno fijo, sino en un entorno que cambia en respuesta a mis acciones.

El objeto de este capítulo es analizar este tipo de situaciones y ver si el concepto de racionalidad que hemos empleado hasta ahora es aplicable en ellas o si, por el contrario, debemos modificarlo. La teoría que estudia la conducta racional en situaciones en las que se produce una interacción entre varios agentes racionales es la Teoría de juegos, donde "juego" es el nombre que se aplica a tales situaciones. La teoría de juegos, que debe su nombre a la utilización de juegos de salón como modelo para analizar las situaciones de interacción, realiza determinados supuestos tanto respecto a la situación de interacción como en relación a los agentes que toman parte en ella. En lo que respecta a la situación, la teoría de juegos asume que está definida por lo que podemos llamar "reglas del juego". Estas reglas especifican

- las acciones que cada jugador puede realizar en cada momento del juego

- la cantidad de información disponible para cada jugador

- las consecuencias que las acciones de cada jugador tendrán tanto para él como para el resto de los jugadores.

En relación con los jugadores, la teoría de juegos asume que

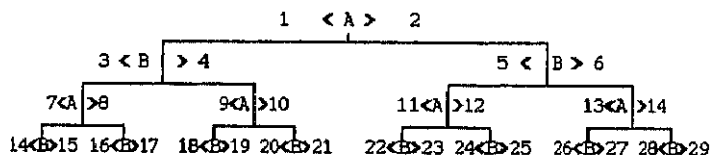
- cada jugador tiene unas determinadas preferencias respecto a cada uno de los posibles resultados del juego y que estas preferencias forman una ordenación

- cada jugador intenta por medio de sus acciones maximizar su utilidad. Es decir, la teoría supone que los participantes en la situación de interacción son individuos racionales.

Una situación de juego determinada puede ser representada o bien del modo conocido como "forma extensiva" o bien en su "forma normal". Básicamente, la diferencia entre ambos modos de representación reside en que la forma extensiva supone que los jugadores realizan una serie determinada de acciones o jugadas alternativamente, e.d., de un modo no simultáneo, mientras que la forma normal supone que cada jugador decide antes de empezar el juego qué es lo que va a hacer en cada una de las posibles situaciones que pueden surgir en el transcurso del juego.

La forma extensiva intenta representar de la forma más intuitiva posible lo que sucede realmente en un juego. Pensemos por ejemplo en una partida de póquer y supongamos, para simplificar, que participan en ella dos jugadores, a los que llamaremos A y B. Una vez que se han repartido

las cartas, uno de ellos, digamos A, tiene que decidir si quiere descartarse y, de ser así, de cuantas cartas (las reglas del juego determinan cuál de los dos debe ser el primero en decidir). El siguiente paso le corresponde al otro jugador, B, quien a su vez tendrá que decidir respecto al descarte. El siguiente paso le corresponde de nuevo a A, quien debe o bien realizar una apuesta determinada o bien "pasar" y cederle la palabra a B y así sucesivamente hasta el final de la partida. La forma extensiva representa este carácter progresivo del juego mediante árboles. Un árbol que representara la partida de póquer sería



Sin embargo, a pesar de su carácter más intuitivo, la forma extensiva de representación tiene el inconveniente de resultar extremadamente engorrosa, pues en juegos mínimamente complejos, ya sea por el número de jugadas, por el número de alternativas en cada jugada o por el número de jugadores, requiere la utilización de unos árboles excesivamente complicados. Por ello, utilizaré como método de representación la forma normal¹. Como dijimos, este modo de representación supone que cada jugador decide de antemano qué es lo que va a hacer en cada una de las posibles situaciones. Es decir, es como si los jugadores decidieran sus movimientos simultáneamente. El conjunto de acciones que un jugador realizará en cada una de las posibles situaciones es su **estrategia**.

La representación gráfica de la forma normal de un juego consiste en una matriz de n entradas, una por cada jugador, en la que se especifican

¹ Por otra parte, la forma normal de representación es la habitual en la discusión de los problemas que aquí nos interesan. En efecto, como veremos más adelante, los problemas más interesantes surgen cuando los jugadores no pueden comunicarse y tienen que decidir qué hacer sin saber lo que los otros van a hacer, o bien cuando cada la situación se define de tal modo que la decisión de cada uno es invariable respecto a lo que hagan los demás.

1. las diferentes estrategias alternativas entre las que puede elegir cada jugador
2. las posibles combinaciones de estrategias, una por cada jugador, y
3. los resultados que cada jugador obtendrá de cada combinación de estrategias.

Por ejemplo, una representación de un juego de dos personas sería la matriz de doble entrada

	B_1	B_2
A_1	(0,0)	(3,4)
A_2	(4,3)	(1,2)

donde las estrategias del jugador 1 son el conjunto A_1, A_2, \dots, A_n , (con sólo dos elementos en nuestro ejemplo) y se corresponden con las filas de la matriz y las estrategias del jugador 2 son el conjunto B_1, B_2, \dots, B_n (también sólo dos en nuestro ejemplo) y se corresponden con las columnas de la matriz. Los números que aparecen entre paréntesis representan el valor de utilidad que cada jugador obtiene de la situación resultante del uso de un determinado par de estrategias, una por cada jugador. Así por ejemplo, si el jugador 1 elige la estrategia A_1 y el jugador 2 la estrategia B_2 , la situación resultante tendrá un valor 3 para el jugador 1 y un valor 4 para el jugador 2.

En una situación de elección en la que el resultado depende de la actuación de más de un agente, las participantes se comportarían de un modo irracional si ignoraran esta característica de la situación. Puesto que las acciones de los demás agentes resultan determinantes para la obtención de un determinado resultado, ignorar este hecho significa ignorar parte de la información relevante que debe ser utilizada para decidir un curso de acción. El agente que ignorara esta información se pondría a sí mismo en la situación de tener que elegir en condiciones poco favorables, con un conocimiento deficiente de la situación y, por tanto, con menores probabilidades de elegir una acción que conduzca al resultado deseado. Por consiguiente, el agente que obrara así, obraría irracionalmente.

Este hecho queda bien expresado utilizando la terminología introducida por Elster, quien distingue entre racionalidad paramétrica y racionalidad estratégica. La racionalidad paramétrica es la que toma el entorno como fijo, donde el único factor variable es el comportamiento del agente. Por el contrario, la racionalidad estratégica toma en cuenta el hecho de que en la situación intervienen otros agentes racionales. Podemos decir entonces que utilizar un modelo paramétrico de racionalidad en situaciones de naturaleza estratégica resulta irracional.

El concepto de racionalidad definido en los capítulos anteriores es un concepto paramétrico de racionalidad. Al tratar de la elección racional en contextos de naturaleza estratégica debemos acudir a un concepto de racionalidad estratégico. Puesto que queremos saber cuál es ese concepto, debemos ver si en estas situaciones tiene aplicación la identificación de acción racional con la maximización de la utilidad esperada. Es decir, tenemos que saber si el uso de la racionalidad estratégica conduce al mismo concepto de acción racional que el uso de la racionalidad paramétrica o si, por el contrario, este concepto debe ser modificado.

Una situación de juego es una situación de interacción. En estas situaciones, un agente estratégicamente racional toma en cuenta este hecho, lo que significa fundamentalmente tomar en cuenta la presencia de otros agente racionales. En tanto que agentes racionales, los individuos con los que se interactúa tendrán un conjunto de preferencias que cumplen las condiciones de racionalidad definidas en capítulos anteriores, de tal

modo que será posible definir una función de utilidad para cada uno de ellos. Dadas estas funciones, la conducta de cada individuo estará dirigida a la maximización de la utilidad.

En una situación de interacción, el agente estratégicamente racional decidirá su conducta no sólo a partir de su propia función de utilidad, sino también a partir de las funciones de utilidad de los individuos con los que interactúa. Es decir, el agente estratégicamente racional tomará en cuenta las preferencias de los otros agentes, las acciones alternativas entre los que los demás pueden elegir y los posibles resultados de esas acciones. Y puesto que la racionalidad estratégica no sólo exige que consideremos al resto de los agentes como racionales, sino también como estratégicamente racionales, cada agente deberá tomar en cuenta que el resto de los agentes decidirá su acción según sus expectativas respecto a las acciones de los demás. Es decir, el agente estratégicamente racional supondrá que interactúa con otros individuos cuyo razonamiento práctico es paralelo al suyo.

Por todo esto, una situación de interacción ideal es aquella en la que cada uno de los agentes tiene pleno conocimiento de todos los parámetros que definen la situación, entre los que se encuentran las funciones de utilidad de los participantes y las posibles acciones que se presentan como alternativas a cada uno.

Un juego en el que todos los participantes tienen esa información es un juego de información completa. La teoría de juegos se ha ocupado con preferencia de este tipo de juegos, tratando de definir la conducta racional en situaciones ideales.

Las situaciones de interacción reales suelen separarse del caso ideal. Normalmente, nos encontramos con casos de juegos de información incompleta, en los cuales no todos los jugadores (y más habitualmente ninguno) cuentan con esa información acerca de los demás jugadores. Para tratar con este tipo de situaciones se ha diseñado un método que reduce un juego de información incompleta a uno de información completa. Este método consiste básicamente en representar la información incompleta acerca de determinados parámetros del juego por parte de los jugadores como si se tratara de información imperfecta, es decir, como si lo que sucediera es que los jugadores desconocen algunos de los movimientos e-

fectuados en el juego con anterioridad a que ellos tengan que tomar su decisión.

Para cumplir nuestro objetivo actual, a saber, analizar si en situaciones de interacción la acción racional puede seguir siendo identificada con la maximización de la utilidad, nos basta con tratar el caso ideal representado por los juegos de información completa. Por lo tanto, nos limitaremos por el momento al análisis de este tipo de juegos.

Hemos dicho que la racionalidad estratégica exige que la acción realizada por cada jugador sea adecuada a la acción que se espera que el resto de los jugadores, que se suponen racionales, realicen. Gauthier explicita este requisito en tres condiciones (Gauthier, 86, p. 61):

A: la elección de cada jugador debe ser la respuesta racional a las elecciones que espera que realicen los otros.

B: cada persona debe esperar que la elección de cada uno de los demás satisfaga la condición A.

C: cada jugador debe creer que su elección y sus expectativas se reflejan en las expectativas de cada uno de los demás.

Puesto que queremos saber si una acción que cumpla las condiciones de racionalidad estratégica puede ser identificada con una acción maximizadora de la utilidad, de tal modo que el concepto de acción racional en situaciones estratégicas sea el mismo que en condiciones paramétricas, debemos investigar la posibilidad de sustituir la expresión "respuesta racional" por "respuesta maximizadora de la utilidad" en la condición A.

Una vez que efectuamos esta sustitución, la condición A exige que la estrategia utilizada por cada jugador sea una mejor respuesta. Una estrategia utilizada por un determinado jugador es una mejor respuesta a las estrategias de los demás jugadores si maximiza la utilidad de este jugador en tanto que las estrategias de los demás jugadores se mantengan. Por tanto, las tres condiciones de Gauthier equivalen a exigir que el resultado del juego sea un resultado de equilibrio. Un resultado es un resultado de equilibrio si es el resultado de una combinación de estrategias, una de cada jugador, tales que cada una es la mejor respuesta a las estrategias de los demás jugadores.

En lo que resta de este capítulo, analizaremos distintas situaciones de interacción, con el objetivo de ver si estas exigencias, que sur-

gen de la identificación de la acción racional con la acción maximizadora de la utilidad, pueden ser cumplidas.

Puesto que en una situación de juego el agente tiene que interactuar con otros agentes que tienen sus propios intereses, un dato importante para la definición de un juego es el tipo de relación que los intereses de los distintos agentes guardan entre sí. Por ello, una clasificación fundamental que establece la teoría de juegos obedece a este dato. Según esta clasificación, los juegos se dividen en tres grupos: juegos con intereses idénticos, juegos con intereses opuestos y juegos con intereses mixtos.

Los juegos con intereses idénticos son aquellos en los que los jugadores tienen las mismas preferencias por los distintos resultados posibles del juego, de modo que si para el jugador 1 xPy , entonces para el jugador 2 xPy , y si para el jugador 1 xIy , entonces para el jugador 2 xIy .

Un ejemplo de juego con intereses idénticos estaría representado por la siguiente matriz:

	B_1	B_2
A_1	(15,15)	(10,10)
A_2	(0,0)	(5,5)

Ejemplo 1

En este juego participan dos jugadores, el jugador 1 y el jugador 2, cada uno de los cuales puede escoger una de dos acciones posibles, siendo las acciones de 1 A_1 y A_2 , y las de 2 B_1 y B_2 .

Lo característico de este tipo de juegos es que cada jugador gana exactamente lo mismo que el otro. Por ello, supuesto que ambos jugadores conocen la matriz que representa el juego, el jugador 1 seguirá la estra-

tegia A_1 y el jugador 2 la estrategia B_1 , ya que cualquier otra combinación de estrategias conduce a un resultado peor para ambos².

El resultado de la combinación de estrategias (A_1, B_1) es un resultado de equilibrio. Por tanto, si ambos jugadores son estratégicamente racionales, el resultado del juego será $(15, 15)$, ya que este resultado cumple las condiciones que hemos establecido para que una acción resulte estratégicamente racional.

En este tipo de juegos es aplicable el concepto de acción racional que hemos utilizado hasta ahora. Como hemos dicho, ese concepto resulta aplicable si en la condición A (p.56) puede decirse "respuesta maximizadora" en lugar de "respuesta racional". En este caso, puede realizarse la sustitución sin problemas, ya que el resultado de equilibrio es el mejor para ambos jugadores.

Los juegos de intereses opuestos representan el caso contrario. Lo característico de estos juegos es que las preferencias de los participantes sobre los posibles resultados del juego son exactamente opuestas. Es decir, si un jugador prefiere X a Y , entonces el otro prefiere Y a X , y si un jugador es indiferente entre x e y , entonces el otro también lo es. De igual modo, sus preferencias sobre loterías serán exactamente opuestas³.

Sabiendo esto sobre las preferencias de los dos jugadores, puede establecerse una medida de utilidad cardinal tal que si X tiene la utili-

² En efecto, el jugador 1, supuesto que es racional, elegirá la estrategia que resulte ser la mejor respuesta a la estrategia que espera que utilice el jugador 2, y este por su parte hará lo mismo. Puesto que sus intereses son idénticos, cada uno sabe que el otro elegirá la estrategia que contribuye a la consecución del resultado que es mejor para ambos.

³ Supongamos, por ejemplo, que el juego tiene cuatro resultados posibles, M, X, Y, Z , y que el jugador 1 los prefiere en el orden Z, Y, M, X , mientras que el jugador 2 los prefiere en el orden X, M, Y, Z . Es decir, para el jugador 1 estos resultados tendrán una utilidad tal que $U(z) > U(y) > U(m) > U(x)$, y para el jugador 2 $U(x) > U(m) > U(y) > U(z)$.

Supongamos ahora que a estos jugadores se les da a elegir entre las dos loterías siguientes:
 L_1 , que le da la probabilidad $2/3$ de obtener Y y la probabilidad $1/3$ de obtener M .
 L_2 , que le da la probabilidad $1/3$ de obtener Y y la probabilidad $2/3$ de obtener M .
 Si calculamos la utilidad de estas loterías para cada uno de los jugadores de la manera habitual, veremos que la utilidad de L_1 para el jugador 1 es de $2/3$ y para el jugador 2 de $1/3$, mientras que la utilidad de L_2 para 1 es de $1/3$ y para 2 de $2/3$. Por tanto, para el jugador 1 $L_1 \succ L_2$ y para el jugador 2 $L_2 \succ L_1$. Es decir, los jugadores tendrán preferencias opuestas por las loterías sobre los resultados, supuesto que sus preferencias por los resultados son opuestas.

dad a para el jugador 1, entonces X tendrá la utilidad b para el jugador 2, donde $b = -a$. Es decir, se pueden representar las utilidades del jugador 1 con números positivos y las de el jugador 2 con números negativos, de modo que todo resultado posible del juego tendrá un pago (a_1, b_1) , donde $a_1 + b_1 = 0^*$.

Debido a que los juegos de intereses opuestos pueden representarse de este modo, es habitual referirse a ellos con la denominación de juegos de suma cero. Nosotros utilizaremos estos nombres indistintamente, pero antes conviene hacer una aclaración respecto a la denominación de "juegos de suma cero".

Al decir que un juego es de suma cero no quiere decirse con ello que en estos juegos un jugador gane en sentido literal lo que otro pierde. Por ejemplo, en una apuesta entre dos personas una gana exactamente el dinero que pierde la otra, es decir, el perdedor paga al ganador determinada suma de dinero o cualquier otra cosa. Desde luego, esto es un caso de juego de intereses opuestos, pero hay otros casos de este tipo de juegos en los que esto no sucede. Un ejemplo de estos otros casos es el ofrecido por Luce y Raiffa (57, p.57), en el que el perdedor es asesinado y se mutila al ganador. Por tanto, al decir que un juego es de suma cero lo único que quiere decirse es que puede ser representado de ese modo.

Para ver cual es la conducta racional en este tipo de juegos analizaremos un ejemplo utilizado por Harsanyi (79, p.149). Este ejemplo esta representado por la siguiente matriz de pagos:

* Esto puede hacerse porque, como dijimos anteriormente, las unidades que se utilizan para medir la utilidad cardinal de una alternativa para un agente son arbitrarias. El método habitual consiste en asignar al jugador 2 los valores 0 y -1 para sus alternativas más y menos preferidas respectivamente. En nuestro ejemplo, si para el jugador 2: $1=0$ y $2=-1$, ofreciéndole las loterías anteriores tendremos que L_1 tiene para el la utilidad $-2/3$ y L_2 la utilidad $-1/3$. Podemos entender ahora que estas loterías se utilizan para asignar valores cardinales a las alternativas de la manera usual, de modo que el jugador 1 es indiferente entre obtener Y con certeza y L_1 , y entre la certeza de W y L_2 , con lo cual sabemos que para el $Y=2/3$ y $W=1/3$. Del mismo modo, 2 es indiferente entre Y y L_1 y entre W y L_2 , con lo que para el $Y=-2/3$ y $W=-1/3$.

	B ₁	B ₂	B ₃
A ₁	(15,-15)	(10,-10)	(18,-18)
A ₂	(9,-9)	(3,-3)	(11,-11)
A ₃	(20,-20)	(6,-6)	(7,-7)

Ejemplo 2.

En esta matriz aparece un resultado, (10,-10), correspondiente a la combinación de estrategias (A₁,B₂) que tiene la característica de ser al mismo tiempo la ganancia mínima de 1 con la estrategia A₁ y la pérdida máxima de 2 con la estrategia B₂. Un resultado con estas características es conocido como punto de encabalgadura (*saddle point*) y puede mostrarse que si ambos jugadores se comportan de un modo racional este será el resultado del juego. En efecto, 10 representa para 1 el peor resultado posible si elige la estrategia A₁, mientras que si siguiera cualquier otra estrategia el peor resultado posible sería peor que 10. Si por ejemplo eligiera la estrategia A₂, el peor resultado posible sería 3 y si siguiera A₃ sería 6. Ciertamente, la estrategia A₁ podría conducir a un resultado que es para A el mejor de todos, a saber, 20. A₃ conduciría a este resultado en el caso de que 2 utilizara la estrategia B₁. Pero, naturalmente, si 2 es un jugador racional nunca utilizará esa estrategia, ya que esto le conduciría al peor resultado posible para él, supuesto que el juego es de intereses opuestos. 2 seguirá la estrategia que, en vista de lo que se espera que 1 haga, le ocasione el mejor resultado posible. Por lo tanto, si 2 actúa racionalmente, nunca conseguirá un resultado peor que -10, ya que utilizando B₂ se asegura este mínimo. Y tampoco conseguirá un resultado mejor, porque si 1 juega racionalmente se asegura una ganancia mínima de 10 utilizando A₁. Habitualmente se conoce con el nombre de nivel de seguridad a la menor utilidad derivable para un jugador del uso de una determinada estrategia. Por tanto, diremos que A₁ y B₁ maximizan el nivel de seguridad de ambos jugadores. Además, A₁ es la mejor respuesta de 1 a la estrategia que espera que 2 utilice. Por su parte,

B_2 es la mejor respuesta de 2 a la estrategia que espera que escoja 1. Por consiguiente, el punto de encabalgadura es un punto de equilibrio.

Puesto que el resultado de la combinación de estrategias (A_1, B_1) es un resultado de equilibrio, los jugadores, al elegir estas estrategias, estarán jugando de un modo estratégicamente racional. En efecto, es fácil comprobar que la elección de estas estrategias cumple las tres condiciones mencionadas en la página 56³.

³ Sin embargo, no todos los juegos de suma cero tienen un resultado que cumpla estas condiciones. Pensemos por ejemplo en un juego con la siguiente matriz de pagos:

	B_1	B_2
A_1	(3, -3)	(1, -1)
A_2	(2, -2)	(4, -4)

Ejemplo 3.

En este juego, A_2 es la estrategia cuyo uso maximiza el nivel de seguridad de 1, mientras que por su parte 2 maximizaría su nivel de seguridad utilizando B_1 . Pero si el jugador 1 espera que 2 utilice B_1 , entonces su mejor respuesta no es A_2 sino A_1 , y si 2 espera que 1 utilice A_1 , entonces su mejor respuesta no es B_1 sino B_2 . Podemos ver entonces que en este juego no existe ningún resultado de equilibrio.

Esto significa que ninguna de las acciones que pueden realizar los participantes en este juego cumple con las condiciones de racionalidad estratégica. Supongamos por ejemplo que el jugador 1 utiliza A_1 . A_1 es la mejor respuesta de 1 si espera que 2 utilice B_1 , con la cual se cumpliría la condición A. Ahora bien, por la condición B, 1 debe esperar que 2 realice B_1 sólo si espera que 2 espera que el va a utilizar A_2 . Sin embargo, puesto que 1 hace A_1 y no A_2 , entonces no se cumple la condición C, que exige que cada jugador sin esperar que sus acciones y expectativas se reflejen en las expectativas de los demás. El mismo razonamiento puede seguirse para el caso de que 1 realice A_2 y para el caso de las dos posibles acciones de 2.

En este juego, por tanto, ningún jugador sabe lo que tiene que hacer, pues no sabe que es lo que hará el otro, y, además, cada jugador sabe que el otro se encuentra en su misma situación.

En esta situación ninguno de los dos jugadores sabe cuál de sus posibles acciones tendrá un mejor resultado. En contextos de naturaleza paramétrica encontramos una situación análoga a esta en las situaciones de no certidumbre. Sin embargo, en situaciones de interacción, la solución no puede ser la misma. El motivo es que ningún agente puede asignar probabilidades a las acciones alternativas de los demás agentes sin tener en cuenta que los otros decidirán su acción dependiendo de lo que esperen que los otros hagan. Por lo tanto, lo único que cada agente puede hacer es asignar una determinada probabilidad a sus propias acciones.

Volviendo a nuestro ejemplo, es de esperar que el jugador 1 sea indiferente entre sus acciones alternativas, de tal modo que este dispuesto a decidir su curso de acción jugando a cara o cruz. Esto sucede porque este método de elección tiene mejores resultados que elegir definitivamente una acción determinada. Es decir, si 1 juega a cara o cruz, entonces, en el caso de que 2 utilice B_1 , 1 tendrá la probabilidad 1/2 de recibir 3 y la misma probabilidad de recibir 2, e.d., al decidir su acción a cara o cruz obtendrá una utilidad esperada de 2,5, lo cual es preferible a la certeza de obtener 2. Y en el caso de que 2 utilice B_2 , el jugador tendrá la misma probabilidad de obtener 1 que de obtener 4, e.d., su utilidad esperada será de nuevo 2,5, lo cual vuelve a ser preferible a la certeza de 2 (recuérdese

A pesar de que los tanto los juegos de intereses opuestos como los juegos de intereses idénticos requieren la aplicación de un razonamiento estratégico, la identificación de la racionalidad práctica con la maximización de la utilidad sigue siendo posible. El motivo es el siguiente. En el caso de los juegos con intereses idénticos los jugadores utilizan la estrategia que llamaremos "cooperativa", e.d., la estrategia que ayuda a el otro jugador a conseguir el máximo de utilidad posible, porque cada jugador se beneficia en la misma medida en que se beneficia el otro. De este modo, ambos jugadores consiguen el máximo beneficio.

Por otro lado, en los juegos de intereses opuestos ninguno de los jugadores puede cooperar con el otro en modo alguno. Puesto que cada uno ganará en la medida en que el otro pierda, cada uno estará interesado en que el beneficio del otro jugador sea el mínimo posible. En estos juegos, cada uno seguirá la estrategia cuya utilidad sea mayor y esto conduce a un resultado que es el mejor posible, e.d., no hay ningún otro resultado que sea mejor para ambos a la vez, pues todo otro resultado que rinda una mayor utilidad a uno de los jugadores tiene como contrapartida una menor utilidad para el otro.

que 2 es su nivel de seguridad si utiliza una estrategia definida).

En general, cada uno de los jugadores en este tipo de situaciones utilizará una estrategia mixta, e.d., asignar una determinada probabilidad a cada una de sus acciones alternativas. La estrategia mixta deberá cumplir dos condiciones:

1. el valor de las probabilidades asignadas a las acciones alternativas sumará 1
2. la utilidad esperada de la estrategia mixta será tal que el agente será indiferente ante cualquier acción del otro jugador.

Volviendo a nuestro ejemplo, los valores de probabilidad p_1 y p_2 para el jugador 1 deberán cumplir:

$$1. p_1 + p_2 = 1$$

$$2. 3p_1 + 2p_2 = 1p_1 + 4p_2$$

y para 2

$$1. p_3 + p_4 = 1$$

$$2. p_1(-3) + p_3(-1) = p_2(-2) + p_4(-1)$$

es decir, en nuestro ejemplo el jugador 1 utilizará la estrategia mixta $(1/2A_1, 1/2A_2)$, y el jugador 2 utilizará la estrategia mixta $(3/4B_1, 1/4B_2)$. La utilización de estas estrategias proporciona a 1 la utilidad esperada 2,5 y a 2 la de -2,5. Para 1, 2,5 es preferible a 2 que es su ganancia mínima máxima con estrategias puras, y para 2 la utilidad -2,5 es preferible a -3, que es su pérdida máxima mínima si utiliza estrategias puras. Es decir, el uso de esas estrategias mixtas eleva el nivel de seguridad de ambos jugadores.

Puede comprobarse fácilmente que el uso de estas estrategias mixtas conduce a un resultado de equilibrio. Es decir, la elección de estas estrategias mixtas cumple con la tres condiciones de racionalidad estratégica.

Vemos por tanto que estos dos tipos de juego no representan ningún problema. Sin embargo, las situaciones de interacción reales no suelen responder a ninguno de estos dos modelos, sino que más bien deben ser representadas como juegos de intereses mixtos. Y es precisamente en este tipo de juegos en el que parecen surgir los problemas.

Los juegos de intereses mixtos son aquellos en los que los intereses de los participantes son en algunos casos similares y en otros opuestos. Es decir, la ordenación que un jugador establece entre los distintos resultados posibles del juego según sus preferencias coloca a algunos de los resultados en el mismo puesto que ocupan en la ordenación del otro jugador, mientras el puesto en el que aparecen otros resultados es el inverso al que ocupan en la ordenación del otro jugador.

En los juegos de intereses mixtos pueden aparecer dos tipos de problemas:

1. puede suceder que el juego tenga dos resultados de equilibrio tales que ambos son óptimos y que los jugadores tienen preferencias opuestas respecto a ellos

2. puede suceder que el juego tenga un resultado de equilibrio no óptimo.

Estudiaremos ambos problemas por separado y en ese orden.

1. Imaginemos un juego representado por la siguiente matriz de pagos:

	B_1	B_2
A_1	(2,1)	(-1,-1)
A_2	(-1,-1)	(1,2)

Ejemplo 4

Este juego tiene dos resultados de equilibrio, (2,1) y (1,2), tales que ambos son óptimos. Se dice que un resultado es óptimo si no es posible encontrar otro resultado que sea mejor para algún jugador sin que sea peor para otro. El problema surge porque, puesto que los dos resultados de equilibrio rinden una utilidad distinta a cada uno de los jugadores, las preferencias de estos con respecto a los resultados de equilibrio son

inversas. Es decir, la ordenación del jugador 1 sobre el conjunto de resultados del juego sería wPz , zPx , xRy , mientras que la del jugador 2 sería zPw , wPx , xRy , donde $w = (2,1)$, $x = (-1,-1)$, $y = (-1,-1)$ y $z = (1,2)$.

Puede mostrarse que si los jugadores tienen que decidir su estrategia independientemente, el resultado de este juego no será ninguno de los dos resultados de equilibrio (demostración en Luce y Raiffa, 57.p.9-2). Es decir, este tipo de situaciones plantean un problema de coordinación. Si a los jugadores se les permite comunicarse y decidir lo que van a hacer de una manera conjunta, tratarán de ponerse de acuerdo para que se realice uno de los dos resultados de equilibrio. Puesto que sus preferencias sobre estos resultados son opuestas, pero ambos están interesadas en que el juego se resuelva de una de las dos maneras, pueden seguir la táctica de decidir a cara o cruz la estrategia que van a utilizar. Ambos estarían de acuerdo en decidir la cuestión de este modo, ya que este les proporciona a cada uno una utilidad de $3/2$.

Supongamos que a los dos jugadores se les permite coordinar sus acciones de este modo y que, tras tirar la moneda, ambos quedan de acuerdo en utilizar A_1 y B_1 respectivamente. Ninguno de ellos estaría interesado en romper este acuerdo. En efecto, puesto que el jugador 1 está interesado en que sea ese precisamente el resultado del juego, no es previsible que rompa el acuerdo. Y una ruptura unilateral por parte del jugador 2 conduciría a un resultado que ninguno de los dos quiere. Nadie gana nada rompiendo el acuerdo.

Este tipo de casos no plantean ningún problema serio. En primer lugar, porque lo único que sucede en estos casos es que es necesaria la coordinación para llegar a un acuerdo pero, una vez que este se logra, resulta sumamente estable. Por otro lado, no es habitual que los jugadores no puedan comunicarse, de modo que los casos en los que el juego no tiene solución no se dan en realidad. En segundo lugar, la solución al problema de coordinación es tan intuitiva que no merece la pena detenerse a defenderla. En efecto, en todos los casos de estas características en los que es preciso tomar una decisión conjunta es práctica habitual decidir la cuestión a cara o cruz.

2. Mucho más interesantes resultan este otro tipo de casos, a cuyo estudio se ha dedicado gran parte de la literatura especializada.

Un ejemplo de este tipo de situaciones sería el siguiente:

	B_1	B_2
A_1	(2,2)	(0,3)
A_2	(3,0)	(1,1)

Ejemplo 5

En este juego cada uno de los participantes tiene dos estrategias entre las que tiene que elegir. Llamaremos a las estrategias A_1 y B_1 estrategias cooperativas, ya que el uso de estas estrategias beneficiaría al contrario, y a las estrategias A_2 y B_2 estrategias no cooperativas, pues su uso perjudica al otro jugador.

En este caso, el único par de estrategias en equilibrio es el par (A_2, B_2) . Más aun, A_2 y B_2 son estrategias dominantes. Es decir, para el jugador 1 es mejor utilizar A_2 sea lo que sea lo que el otro haga. Si 2 utiliza B_1 , entonces lo mejor que 1 puede hacer es utilizar A_1 , pues de este modo su utilidad sería de 3 en vez de ser de 2 y si el jugador 2 utiliza B_2 , de nuevo lo mejor que 1 puede hacer es utilizar A_1 , pues esto le produciría un beneficio de 1 en lugar del 0, que sería lo que conseguiría si obrará de otro modo. Un razonamiento paralelo puede hacerse para el jugador 2.

Por consiguiente, cada uno de los jugadores se comportará racionalmente si utiliza esas estrategias. Sin embargo, su uso conduce a un resultado que no es óptimo. El juego tiene otro resultado que es mejor para ambos jugadores.

Puede pensarse que en este caso nos encontramos con otro problema de coordinación. Es fácil comprobar que esto no es así.

Una de las diferencias fundamentales entre este tipo de juegos y los juegos de suma cero es que en este caso los participantes pueden optar por cooperar. No solamente la cooperación es una alternativa, sino que es una alternativa que conduciría a un resultado óptimo. Ambos jugadores saben que si no cooperan ambos, el resultado será inferior a lo que podría ser en caso de que cooperasen. De modo que podemos suponer que, si a los participantes se les permite coordinar sus acciones, llegarán a un acuerdo de cooperación. Supongamos que esto es lo que sucede.

Sin embargo, si ambos jugadores son racionales, este acuerdo nunca se cumpliría. Pongámonos en la situación del jugador 1. Este jugador sabe

que, en caso de que el jugador 2 mantenga el acuerdo y utilice B₁, lo mejor que el puede hacer es no mantener el acuerdo y utilizar A₂. Y, en caso de que el otro jugador rompa el acuerdo, también saldrá mejor parado si utiliza la estrategia no cooperativa.

Es decir, puesto que las estrategias no cooperativas son las estrategias dominantes, ninguno de los dos jugadores tendrá ningún motivo para mantener el acuerdo. Naturalmente, ambos jugadores saben que su razonamiento será duplicado por el otro jugador, a resultas de lo cual el resultado del juego será (1,1). Podemos suponer que ninguno de ellos desea romper el acuerdo siempre y cuando no lo haga el otro. Pero, a pesar de que ambos han decidido conjuntamente utilizar la estrategia cooperativa, ninguno de ellos tiene ningún motivo para suponer que el otro va a mantener el acuerdo. Es más, cada uno sabe que el otro está fuertemente tentado (tanto como él mismo) para utilizar la estrategia no cooperativa. De modo que ambos decidirán cubrirse las espaldas utilizando la estrategia no cooperativa.

Por tanto, el problema que plantean este tipo de juegos es distinto al que surge en el caso del problema de coordinación. Vimos que en esos casos la consecución de un resultado óptimo (e.d., de uno de los dos existentes en el juego) era individualmente inaccesible. Sólo coordinando sus acciones podían los jugadores acceder a este resultado. Pero una vez que se producía la coordinación, el acuerdo conseguido era sumamente estable. Sin embargo, en el caso presente el resultado óptimo no sólo es individualmente inaccesible, sino que además es inestable: a ninguno le interesa mantenerlo, de modo que ambos tienen buenas razones para suponer que el otro actuará de un modo no cooperativo.

Este tipo de casos, habitualmente conocidos como el Dilema del prisionero, plantean un serio problema a la identificación de la racionalidad práctica con la maximización de la utilidad. Cuando en situaciones de este tipo los jugadores actúan de un modo racional, e.d., intentando maximizar la utilidad, la consecuencia es que el resultado del juego no es óptimo. Y los jugadores consiguen este resultado no óptimo precisamente por que son jugadores racionales.

El problema planteado por el dilema del prisionero no es un problema de coordinación, sino de cooperación. Ambos jugadores conseguirían un resultado mejor si cooperasen. Pero para esto no es suficiente que los

participantes en las situación lleguen a un acuerdo. Además, es necesario que la situación en la que se desarrolla el juego tenga unas características determinadas. Esta situación debe ser tal que cada participante pueda estar completamente seguro de que el otro cooperará. Y esta seguridad sólo puede lograrse si cada uno sabe que el otro no estará tentado a actuar de un modo no cooperativo.

En los capítulos siguientes estudiaremos con detalle los problemas de cooperación y sus posibles soluciones. Sin embargo, antes de comenzar ese análisis es conveniente recapitular las condiciones que han hecho posible el surgimiento del dilema.

Parfit (86,p.56) resume en tres las condiciones que hacen posible el surgimiento del dilema. Este puede aparecer cuando:

a.-Tenemos una teoría relativa al agente, e.d., una teoría que propone a cada uno de los agentes un fin distinto

b.-la consecución de los fines de cada uno de los agentes depende en parte de lo que los demás hagan

c.- lo que hace cada uno no afecta a lo que hacen los otros.

La teoría sobre la racionalidad práctica que nosotros estamos defendiendo cumple la primera condición. En efecto, esta teoría dice que cada agente se comportará racionalmente si intenta maximizar su utilidad. Por tanto, en una teoría en la que se cumple esa condición es posible el surgimiento del dilema cuando los agentes se encuentran en una situación de interacción (condición b), en la que cada jugador tiene una estrategia dominante (condición c).

Si estas condiciones se cumplen, entonces se darán casos en los cuales si cada uno de los jugadores se comporta de acuerdo con la teoría, e.d., si cada uno hace lo que es mejor para sí mismo, el resultado será peor para ambos que si no lo hicieran.

Este tipo de casos se producirán cuando se dan dos condiciones:

Condición positiva, cuando cada uno puede elegir entre (1) asegurarse a sí mismo el menor de dos beneficios o (2) proporcionar al otro el mayor beneficio.

Condición negativa, cuando las elecciones que realicen los agentes no serán en ningún otro sentido ni mejores ni peores para ninguno.

La condición negativa requiere que el juego que plantea el dilema no se repita de nuevo entre los mismos jugadores. Si los mismos jugado-

res se enfrentaran repetidamente con la misma situación la elección de cada uno podría influir en lo que el otro hace en la siguiente jugada. Es decir, si los mismos jugadores se enfrentan con el mismo dilema H en momentos sucesivos, es previsible que lo que el jugador 1 hace en la instancia H_1 del juego sea tomado en cuenta por el jugador 2 en la instancia H_2 y viceversa. Sin embargo, ha sido suficientemente probado (por ejemplo, Luca y Raiffe, 57, pp. 97 y ss.) que si los participantes en el dilema iterativo saben exactamente cuantas veces va a repetirse el juego, entonces sigue siendo cierto que si los jugadores son racionales siempre seguirán la estrategia no cooperativa.

En las situaciones reales los jugadores no saben cuántas veces van a verse enfrentados con la misma situación. En estas situaciones no se da la condición negativa y, por tanto, aunque en un principio la situación parezca tener la estructura del dilema, e.d., aunque cada instancia del juego tenga la estructura del dilema, es muy probable que a lo largo de las distintas jugadas llegue a surgir la cooperación. Más adelante discutiremos con más detalle este tipo de casos, pero, de momento, podemos afirmar que en estos casos no se presenta el dilema.

Sin embargo, no debe por ello pensarse que el problema planteado por el dilema del prisionero no es más que un problema interno a la teoría de juegos que no tiene ejemplificaciones en la realidad y que ocuparse de ello es un entretenimiento teórico comparable a las aporías de Zenón (uno siempre puede pasar el rato planteando a los amigos el curioso caso de Aquiles y la tortuga, pero, al fin y al cabo, en la carrera real, Aquiles alcanza a la tortuga ¿no?). Porque aunque pueda ser dudoso que en la práctica se den muchos casos reales del dilema en los que participan dos jugadores, se dan multitud de casos en los que el agente se enfrenta a dilemas multipersonales. En estas situaciones el dilema consiste en que si cada uno hace lo que es mejor para sí mismo, entonces el resultado es peor para todos. Uno de los ejemplos más habituales de este tipo de dilemas es el que se conoce con el nombre de "dilema del contribuyente": cada uno se beneficiaría más si no pagase impuestos, pero si ninguno los paga el resultado es peor para todos. Este tipo de dilemas aparecen cuando

Condición positiva, (1) cada uno podría, con algún sacrificio por su parte, proporcionar a otros un beneficio mayor y (2) si cada uno proporciona ese beneficio a los otros, cada uno resultará más beneficiado

-- *Condición negativa*, no hay otros efectos indirectos de las acciones de cada uno que supriman esos efectos directos.

Contrariamente a lo que sucede en los dilemas bipersonales, en estos casos la condición negativa suele mantenerse. El motivo es que es muy improbable que lo que uno hace repercute en lo que hacen los demás.

Parfit afirma que una teoría que cumpla las tres condiciones que posibilitan la aparición de este tipo de dilemas es una teoría directamente auto-frustrante a nivel colectivo. Esta característica queda definida del siguiente modo:

Una teoría T es directamente auto-frustrante a nivel colectivo cuando sucede que, si todos siguen con éxito lo que exige T, entonces y a consecuencia de ello el fin que T propone a cada uno se logra peor que si ninguno hubiera seguido satisfactoriamente T (Parfit, 86, p. 55).

La teoría de la racionalidad práctica que hemos expuesto propone a cada agente un fin que debe cumplir para ser un agente racional, a saber, maximizar la utilidad. Lo que muestra el dilema es que si todos los participantes en el juego intentan maximizar su utilidad, el resultado no es óptimo. En estos casos, si los jugadores no hubieran actuado racionalmente según la teoría, sino que la hubieran desobedecido, el resultado hubiera sido óptimo.

En este punto resulta necesario matizar el problema. Lo que nuestra teoría exige no es que el agente consiga *de hecho* el mejor resultado posible, sino que su conducta este dirigida a conseguir el máximo de utilidad posible que permiten las circunstancias. Dadas las circunstancias que caracterizan el dilema, cada agente se comporta racionalmente si elige la estrategia no cooperativa. Desde el momento en que nuestra teoría se plantea como una teoría acerca de la racionalidad individual, que sea auto-frustrante a nivel colectivo no resulta en principio una objeción.

Sin embargo, el fracaso a nivel colectivo es una objeción en otro sentido. Desde luego, si el juego se desarrolla de un modo no cooperativo, los agentes se comportarán de un modo racional si no cooperan. Pero si es posible transformar el juego no cooperativo en un juego cooperativo, es razonable suponer que nuestra teoría exigiría que los agentes ra-

cionales hicieran lo posible por producir en el juego esta transformación. El motivo es que si el juego se resuelve de un modo cooperativo, cada uno de los agentes conseguirían un resultado mejor. Si, por ejemplo, fuera posible establecer acuerdos vinculantes entre los jugadores, la teoría diría que un jugador racional debe hacer lo posible por establecer tales acuerdos. Algo parecido podría decirse respecto a todos los medios posibles de solución del dilema. Respecto a todos ellos, la teoría exige que intenten hacerse efectivos.

Una vez que estudiemos los distintos modos de solución del dilema, podrá replantearse la cuestión de si las distintas soluciones contradicen o no la teoría y, de ser así, qué implicaciones tiene este hecho para nuestro concepto de racionalidad práctica.

SEGUNDA PARTE

EL PROBLEMA DE LA COOPERACION RACIONAL

El objeto de esta parte del trabajo es analizar las posibles soluciones al problema planteado por el dilema del prisionero. Como se recordará, este surge cuando la aplicación de los criterios de racionalidad individual a determinadas situaciones de interacción conduce a resultados no óptimos. El dilema del prisionero nos muestra un caso en el que aparece un resultado óptimo que es al mismo tiempo individualmente inaccesible e inestable. Todos los participantes en el juego están interesados en la formación de acuerdos cooperativos cuyo resultado sea óptimo. Sin embargo, cada uno de los jugadores estaría fuertemente tentado a incumplir el acuerdo. Por consiguiente, la solución del dilema requiere que los acuerdos vayan apoyados por algún tipo de mecanismo que asegure su cumplimiento.

El problema planteado por el dilema puede generalizarse a todas las situaciones de interacción que hemos caracterizado como juegos de intereses mixtos. Sin embargo, la mayoría de esas situaciones son mucho más complejas. Parte de esa complejidad consiste en el hecho de que en la mayoría de las situaciones en las que sería racional llegar a un acuerdo cooperativo no existe un único resultado óptimo. Por el contrario, los posibles resultados óptimos son múltiples y los jugadores tienen preferencias opuestas respecto a ellos. Por lo tanto, antes de estudiar los distintos modos de asegurar el cumplimiento de los acuerdos, es necesario analizar cuál de entre todos los resultados óptimos posibles es tal que unos agentes racionales harían de él el objeto del acuerdo. El primer capítulo de la segunda parte estará dedicado a este problema, dejando para más adelante la cuestión de los medios de los que se dispone para asegurar el cumplimiento de los acuerdos.

CAPITULO V

ACUERDOS RACIONALES

Las situaciones en las que es racional cooperar son múltiples. Dos hermanos reciben una herencia de un tío de América. Si logran ponerse de acuerdo podrán repartírsela. Si no, los pleitos acabaran con la herencia. Unos naufragos en un isla encuentran unos árboles frutales. Si se ponen de acuerdo en como repartir los frutos, su supervivencia esta asegurada. Si no, la lucha entre ellos los agotará y cada uno impedirá que el otro coma. Unos amigos tienen algo de dinero y de fuerza cada uno. Si logran cooperar, podrán organizar un negocio rentable. En caso contrario, cada uno tendrá que seguir trabajando como asalariado. Un profesor llega a una clase con alumnos especialmente dispuestos a molestar y sin ninguna disposición al estudio. Si consiguen llegar a un acuerdo, el profesor dará sus clases tranquilamente y los alumnos aprenderán. Ninguno de ellos tendrá problemas. Si no lo consiguen, el profesor vera su trabajo convertido en una tortura diaria y los alumnos suspenderán el curso. Todos tendrán un año académico desagradable y lleno de tensiones.

Su característica común es que son situaciones en las que, si el juego se desarrolla de un modo no cooperativo, el resultado de equilibrio no es óptimo. Puesto que en estas situaciones todos los participantes resultarían beneficiados si cooperaran, todos están interesados en la cooperación. Si los jugadores consiguen llegar a un acuerdo todos saldrán ganando.

El objeto de los acuerdos cooperativos es la elección de un conjunto de estrategias, una para cada uno de los participantes en el juego. Es decir, los participantes en el juego deben decidir conjuntamente que es lo que cada uno de ellos va a hacer. La decisión que se tome en cada caso dependerá del resultado que se espera como consecuencia de cada uno de los conjuntos de estrategias. Si los jugadores son racionales, sólo estarán de acuerdo en la elección de un conjunto de estrategias si su resultado es óptimo. Ahora bien, casi todos los casos cuentan con más de un resultado óptimo. En estos casos, el problema consiste en que cada uno de los jugadores tiene preferencias opuestas respecto a los distintos resultados óptimos. Supongamos por ejemplo que dos personas se encuentran en una situación en la cual tiene que repartirse 100 pesetas. Supongamos

también que si no logran ponerse de acuerdo en como van a repartirlas, ninguno conseguirá nada. Por tanto, lo racional para ambos es lograr un acuerdo. Pero hay muchos modos de repartir 100 pesetas entre dos individuos y todas ellas son óptimas. Por ejemplo, tanto la partición (10,90) como la partición (90,10) son óptimas. Sin embargo es evidente que el jugador 1 prefiere la segunda partición mientras que el jugador 2 prefiere la primera¹. Por consiguiente, establecer la condición de que el resultado del acuerdo debe ser óptimo no resulta suficiente para elegir un resultado concreto. Lo que necesitamos es saber cual de estos resultados resultaría elegido por jugadores racionales.

Un modo de averiguar qué características debe tener un resultado óptimo para resultar elegido como objeto del acuerdo cooperativo consiste en preguntarse qué es lo que un agente racional que se encuentre en situación de realizar un acuerdo puede exigir. En tanto que agente racional, su pretensión será que el acuerdo sea tan favorable para él como sea posible. Sin embargo, esta pretensión estará limitada por la propia naturaleza estratégica de la situación. Es decir, el agente racional deberá tener en cuenta que el resultado sólo puede ser conseguido con la cooperación de otros agentes racionales y que por tanto el resultado debe ser no sólo aceptable por él, sino también por todos los demás. Por este motivo, un agente racional exigirá que el resultado del acuerdo sea lo más favorable posible para él sin que por ello deje de ser aceptable por los demás.

El modelo clásico de solución cooperativa a este tipo de situaciones fue presentado por Nash en 1950 (Nash, 50). La idea básica de esta solución es plantear la situación en la que unos agentes racionales tienen que llegar a un acuerdo cooperativo como un juego de intereses opuestos, cuya solución será un resultado de equilibrio. Este planteamiento refleja el hecho de que la solución al problema de cooperación pasa por elegir un miembro de un conjunto de resultados óptimos sobre los cuales los distintos agentes tienen intereses opuestos. Por tanto, Nash plantea

¹ Naturalmente, no es casual que los jugadores tengan preferencias opuestas respecto a los distintos resultados óptimos. Dada la definición de resultado óptimo, esto sucede necesariamente. Puesto que un resultado óptimo es aquel que no puede ser mejor para ningún jugador a menos que sea peor para el otro, si existen más de un resultado óptimo entonces necesariamente si uno de ellos es mejor para un jugador es peor para el otro.

el proceso por el que se logra un acuerdo como un proceso de regateo. Aunque la finalidad de este proceso sea un acuerdo cooperativo, el proceso mismo no es cooperativo sino, por el contrario, estrictamente competitivo, y en el cual cada uno de los agentes intentara maximizar su utilidad².

Un problema de regateo esta definido por las siguientes características:

- un conjunto E de posibles estrategias conjuntas, entre las que los jugadores deberán elegir.

- un conjunto de pares (u,v) , tal que cada miembro de este conjunto se asocia con un miembro del conjunto E , que representan la utilidad que el resultado de la estrategia conjunta tiene para el jugador 1 y para el jugador 2 respectivamente

- un par (u^*,v^*) que representa la utilidad que para los jugadores tiene un miembro especial de E , E^* , que consiste en la estrategia de no cooperación. Dicho de otro modo, (u^*,v^*) representa lo que cada uno de los jugadores conseguiría en el caso de que la negociación no tuviera éxito y el juego se resolviera de una manera no cooperativa. Por este motivo, el punto (u^*,v^*) es conocido como *statu quo*.

En las discusiones del problema de regateo es habitual la representación gráfica del problema. El proceso de representación gráfica es el siguiente:

- las utilidades del jugador 1 se representan en la abscisa y las del jugador 2 en la ordenada

- se sitúan en el plano los distintos pares (u,v) correspondientes a la utilización de estrategias puras. Si la estrategia pura E_1 tiene como resultado el par (u_1,v_1) y la estrategia E_2 el par (u_2,v_2) , entonces los resultados de las posibles estrategias mixtas que resultan de asignar una determinada probabilidad p a E_1 y una probabilidad $1-p$ a E_2 se encuentran representados en el segmento del plano que une (u_1,v_1) y (u_2,v_2) .

- el conjunto de todos los puntos (u,v) correspondientes a los re-

² La solución al problema de regateo planteada inicialmente por Nash esta elaborada para juegos de dos participantes. Esta solución puede ser extendida sin mucha dificultad a juegos multipersonales. Por tanto, y debido a su mayor simplicidad, analizaremos la solución para juegos de dos jugadores.

sultados tanto de las estrategias conjuntas puras como de las mixtas quedara representado por una región R del plano convexa y cerrada.

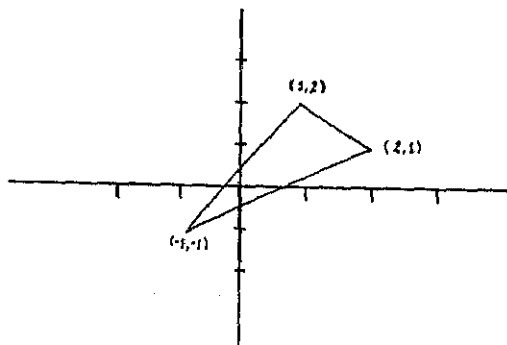
— un problema de regateo queda caracterizado por la fórmula:

$$[R, (u^*, v^*)].$$

Pensemos en el ejemplo 4 del capítulo anterior (p. 63)

	B_1	B_2
A_1	(2,1)	(-1,-1)
A_2	(-1,-1)	(1,2)

La representación gráfica de este caso sería



Una solución al problema de regateo consistirá por tanto en una función que elija para cada caso un punto (u, v) . La función ofrecida por Nash como solución al problema de regateo opera del siguiente modo:

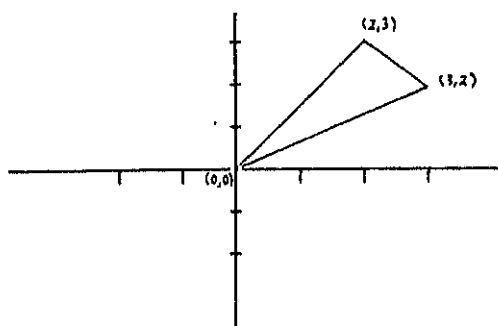
1. Las funciones de utilidad de los jugadores se transforman de tal modo que al *statu quo* se le asigna la utilidad (0,0). Esto significa la transformación de la región R en otra R' .

2. En R' se encontrara un único punto (u_0', v_0') tal que $u_0'v_0'$ es mayor que cualquier otro producto $u'v'$. Este punto será la solución al problema de regateo $[R'(0,0)]$.

3. La solución a $[R(u^*, v^*)]$ se obtiene invirtiendo la transformación de las funciones de utilidad. También puede obtenerse directamente

como el punto (u_0, v_0) tal que $(u_0 - u^*)(v_0 - v^*) = (u - u^*)(v - v^*)$ para todo u y v pertenecientes a R y tales que $u=u^*$ y $v=v^*$.

En nuestro ejemplo, obtendríamos



La solución al problema originario $[R(u^*, v^*)]$ sería el punto $(3/2, 3/2)^3$.

El atractivo de la solución de Nash reside en que satisface una serie de condiciones que parece que deben exigirse una solución aceptable del problema de regateo. Estas condiciones son:

1. Una solución al problema de regateo debe ser invariante respecto a las transformaciones de las funciones de utilidad de los individuos participantes en el regateo. Cumplir con esta condición resulta necesario desde el momento en que la elección de unas determinadas unidades para medir la utilidad de los agentes es arbitraria y que las funciones de utilidad de los individuos son únicas salvo transformación lineal⁴.

2. La solución al problema de regateo debe ser una función que seleccione un resultado óptimo. Es decir, el resultado seleccionado (u, v) debe ser mejor para ambos individuos que el *statu quo* y tal que no haya otro resultado posible que sea mejor para ambos a la vez. Esta condición se limita a reflejar la exigencia de que la solución al problema de regateo

³ Esta solución corresponde a lo dicho anteriormente (cap.4, p.17) sobre este caso. En efecto, la solución $(3/2, 3/2)$ corresponde a tirar un moneda para decidir si se utilizan el par de estrategias (A_1, B_1) o bien el par (A_2, B_2) , cuyos resultados son $(2,1)$ y $(1,2)$ respectivamente. Utilizando esta estrategia mixta conjunta, ambos jugadores obtienen una utilidad esperada de $3/2$.

⁴ Véase Teoría de Juegos, p. 10.

teo sea tal que a los individuos que toman parte en él les resulte ventajoso llevarlo a cabo.

3. La solución debe ser independiente de las alternativas irrelevantes. Esta condición exige que si un problema de regateo se ve ampliado por la introducción de nuevas estrategias posibles, entonces, siempre y cuando el statu quo permanezca igual, la solución a este problema ampliado o bien es la misma que tenía el problema originario o bien es una solución nueva perteneciente a alguna de las nuevas estrategias³.

4. La solución al problema de regateo debe ser tal que, si en un problema de regateo concreto $\{R, (u^*, v^*)\}$ los jugadores desempeñan roles simétricos, entonces su solución (u, v) cumplirá $u = v$. Se dice que los roles de los jugadores son simétricos si sucede que, si i) $u^* = v^*$ y ii) (u, v) es un punto de R , entonces (v, u) también lo es. Esta condición refleja el supuesto de igual racionalidad e igual capacidad de regateo que la teoría supone para los agentes. En efecto, supuesta esta igualdad, y supuesto que no hay ninguna característica del juego que haga las situaciones en las que se encuentran los agentes relevantemente distintas, ninguno de ellos dará su conformidad a un resultado cuya utilidad para él sea inferior a la utilidad que tiene para el contrario.

Estas cuatro condiciones, con la posible excepción de la tercera, son difícilmente discutibles. La solución de Nash no sólo cumple estas condiciones, sino que puede demostrarse que es la única solución que las cumple, e.d., que cualquier otra función distinta que cumpla estas condiciones es equivalente a la función de Nash⁴.

El hecho de cumplir estas condiciones hace que la solución de Nash sea muy intuitiva. Podemos ver esto con un ejemplo. Supongamos que dos personas se encuentran en una situación en la cual tienen oportunidad de repartirse 100 pesetas. Supongamos para simplificar que si no cooperan ninguno conseguirá nada. Por lo tanto, ambos están interesados en establecer un acuerdo que les permita cooperar. Supongamos además que las cien pesetas pueden repartirse de cualquier modo y que la utilidad de los

³ Sin duda, esta es la condición más controvertida. De hecho, buena parte de las críticas a la solución de Nash pasan por una crítica a esta condición. Para una discusión sobre ese punto, ver Luce y Raiffa (57), pp.132 y ss.

⁴ Prueba formal en Luce y Raiffa, 57, pp.127-8.

posibles resultados del acuerdo es lineal con el dinero conseguido. Es decir, los dos jugadores tienen que ponerse de acuerdo en uno de los resultados $\{(0,100), (1,99), (2,98), \dots, (98,2), (99,1), (100,0)\}$.

La solución de Nash sería un punto (u_0, v_0) tal que $u_0 v_0 \geq uv$ para todo u y v pertenecientes a R . En el caso de regateo de nuestro ejemplo el punto (u_0, v_0) sería $(50,50)$.

Esta solución igualitaria tiene un fuerte apoyo intuitivo. En efecto, si dos agentes relevantemente similares deben llegar a un acuerdo sobre como repartirse 100 pesetas, parece que el acuerdo debe consistir en una partición igualitaria. Sin embargo, resulta interesante señalar que este apoyo intuitivo no tiene nada que ver con ningún tipo de intuición moral. La moral no desempeña aquí ningún papel. Y precisamente por esto no llamamos a esta solución "justa" sino "igualitaria", entendiendo este término en sentido puramente descriptivo. Por el contrario, si este resultado resulta acertado intuitivamente es debido a que cumple con las cuatro condiciones mencionadas arriba, especialmente con la cuarta. En efecto, supongamos un acuerdo cercano al de $(50,50)$ pero ligeramente más favorable para uno de los dos jugadores, por ejemplo, $(49.5, 50.5)$. Pero si los roles de los jugadores son simétricos, en especial si los dos pierden lo mismo en caso de que no se llegue a un acuerdo, entonces ¿por qué iba 1 a consentir un acuerdo que es peor para él que para 2? Si lo hiciera no se comportaría como un jugador racional. El jugador 1, supuesta su racionalidad, pretende conseguir el mejor resultado para él, y esta pretensión sólo está limitada por la necesidad de negociar con otro agente racional cuyas pretensiones son iguales a las suyas. Es decir, si bien 1 está interesado en conseguir lo más posible, aún está más interesado en llegar a un acuerdo, de modo que sus pretensiones no pueden ser tan grandes que hagan que el otro jugador no quiera negociar. Pero debido a la simetría de sus posiciones, este mismo razonamiento es duplicado por el otro jugador. De modo que, por una parte, ninguno de ellos accederá a un resultado peor que el igualitario para él y, por otra parte, debido a que el otro está en la misma situación, ninguno conseguirá tampoco un resultado mejor para él que el resultado igualitario.

Sin embargo, aunque la solución de Nash es intuitiva, en sí misma no ofrece ninguna explicación del tipo que nosotros hemos apuntado, e.d., una explicación de por qué unos jugadores racionales deberían estar con-

formes con esa solución.

Harsanyi (Harsanyi, 56) muestra que la solución propuesta por Zeuthen en 1930 es equivalente a la solución de Nash y que tiene la ventaja de ofrecer un "modelo psicológico plausible" de los procesos de regateo reales.

En líneas generales, el modelo de Zeuthen es el siguiente. Al comienzo del proceso de regateo cada una de las partes propone unos términos del acuerdo a la aceptación del otro. Estas propuestas iniciales representan lo que cada uno de los participantes desearía ganar idealmente. Llamemos A_1 a la propuesta del jugador 1 y A_2 a la del jugador 2. $U_1(A_1)$ y $U_2(A_1)$ a la utilidad que representa A_1 para el jugador 1 y el jugador 2 respectivamente y $U_1(A_2)$ y $U_2(A_2)$ a la utilidad de A_2 . Puesto que las propuestas iniciales son *desiderata* de cada uno de los jugadores, $U_1(A_1) > U_1(A_2)$ y $U_2(A_2) > U_2(A_1)$. Una vez que se han hecho las propuestas iniciales, cada una de las partes debe calcular si aceptar la propuesta del otro le producirá una mayor utilidad que insistir en que se apruebe su propia propuesta. Para realizar este cálculo se deben tener en cuenta las probabilidades de que el contrario ceda. Llamemos p_1 a la probabilidad de que el jugador 1 rechace definitivamente la propuesta del jugador 2 y p_2 a la probabilidad respectiva del jugador 2. Si el jugador 1 rechaza la propuesta del contrario, entonces tendrá la probabilidad $1 - p_2$ de recibir $U_1(A_1)$ y la probabilidad p_2 de que la negociación llegue a un punto muerto y no consiga nada. De modo que 1 deberá insistir en su propuesta inicial si ocurre que $(1 - p_2) \cdot U_1(A_1) > U_1(A_2)$ y debe aceptar la propuesta de 2 en caso contrario. Dicho de otro modo, el jugador 1 deberá aceptar la propuesta A_2 si el cociente $(U_1(A_1) - U_1(A_2)) / U_1(A_1) < p_2$. Siguiendo un razonamiento paralelo para el caso del jugador 2, este deberá aceptar A_1 si $(U_2(A_2) - U_2(A_1)) / U_2(A_2) < p_1$.

Estos cocientes expresan el riesgo máximo que cada una de las partes está dispuesta a correr para conseguir que se apruebe su propuesta en vez de la del contrario. Al mismo tiempo, miden la ventaja relativa que A_1 tiene sobre A_2 para el jugador 1 y la que tiene A_2 sobre A_1 para 2 y por consiguiente la determinación con la que cada uno defenderá su propuesta inicial.

A continuación, Zeuthen introduce el supuesto de que una parte debe hacer una concesión si la determinación del contrario es más firme

que la suya propia, e.d., 1 debe conceder si $U_1(A_1) - U_1(A_2) / U_1(A_1) < U_2(A_2) - U_2(A_1) / U_2(A_2)$ y 2 debe conceder en caso de que la inecuación aparezca con el signo opuesto. Hacer una concesión no significa forzosa-mente aceptar la propuesta del contrario. La parte que tiene que ceder puede hacer una nueva propuesta A_3 cuya utilidad sea para él mayor que la de la propuesta del contrario, aunque algo inferior a la de su propia propuesta anterior. De hecho, basta con que esta nueva propuesta sea su- ficiente como para cambiar de signo a la inecuación y hacer ceder al con- trario. De este modo se inicia el proceso de regateo que terminará cuando las sucesivas propuestas de los participantes coincidan en un punto in- termedio⁷.

Harsanyi hace notar que la inecuación de Zeuthen es equivalente a la inecuación $U_1(A_1) \cdot U_2(A_1) < U_1(A_2) \cdot U_2(A_2)$. Es decir, la propuesta de Zeuthen equivale a decir que la parte cuya propuesta tenga asociado un producto menor de utilidades debe hacer una concesión planteando una nue- va propuesta cuyo producto de utilidades sea mayor que el del contrario. A lo largo de la negociación irán surgiendo propuestas cuyo producto de utilidades será cada vez más alto. Si suponemos que en el caso de que la propuesta de ambas partes tenga un producto de utilidades con el mismo valor ambos deben ceder, el proceso de regateo continuara hasta llegar a una propuesta que maximice el producto $U_1 \cdot U_2$, y que será el que establez- ca los términos del acuerdo. Pero como se recordara este punto es preci- samente la solución de Nash al problema del regateo, con lo cual parece demostrado de los planteamientos de Zeuthen y Nash son equivalentes⁸.

⁷ Para un ejemplo detallado del funcionamiento del modelo de Zeuthen, ver Apéndice 6.

⁸ Una vez establecida la equivalencia de ambas soluciones, Harsanyi demuestra que el supuesto de Zeuthen respecto a que parte debe realizar concesiones (a partir de ahora, nos referiremos a este supuesto con el nombre de "principio de concesión de Zeuthen") puede derivarse a partir de unos cuantos postulados muy simples y de carácter muy general. Estos postulados son los siguientes:

I. Simetría. Los agentes que participan en el proceso de regateo siguen las mismas reglas de conducta.

II. Conocimiento perfecto. Cada una de las partes puede hacer una estimación correcta de la probabilidad que existe de que el contrario rechace de modo definitivo una propuesta determinada.

III. Monotonía. La probabilidad de que un jugador rechace una propuesta de la parte contraria es una función monótona no decreciente de la diferencia entre la utilidad que tiene para el su propia propuesta menos la que tiene para el la propuesta del contrario, siempre y cuando el resto de variables permanezcan inalteradas.

IV. Maximización de la utilidad esperada. A menos que ambas partes estén de acuerdo en hacer

En virtud de esta equivalencia, todo lo dicho acerca del carácter intuitivo de la solución de Nash puede aplicarse al principio Zeuthen-Harsanyi. Además, este último tiene la ventaja adicional que reflejar el razonamiento por el que los jugadores llegan a esa solución.

Una vez mostrada la equivalencia de ambos métodos, hablaremos del "procedimiento Nash-Zeuthen-Harsanyi" ("procedimiento N-Z-H" para abreviar), nombre que por otro lado es habitual en la literatura. El resultado de la aplicación de este método puede generalizarse del modo siguiente:

1. Sean U^* y V^* las utilidades respectivas del jugador 1 y del jugador 2 obtenidas en el caso de que no se llegue a un acuerdo y el juego se desarrolle de una manera no cooperativa. El punto (U^*, V^*) representa el *statu quo*. Utilizando el procedimiento N-Z-H, siempre que $U^* = V^*$ el punto de acuerdo será un resultado igualitario, e.d., será un punto (U, V) tal que $U = V$. De nuevo, conviene señalar que este resultado no tiene nada que ver con la aplicación de ningún tipo de criterio de justicia. Todo lo que significa es que, bajo el supuesto de una simetría completa de los puestos ocupados por los participantes en el proceso de regateo, unos jugadores racionales ni aceptarían un resultado no igualitario en su contra ni conseguirían un resultado no igualitario a su favor. Dicho de modo más gráfico, esto significa que si los dos jugadores tienen la misma "fuerza" ambos conseguirán lo mismo en el regateo.

2. En los casos en los que U^* es distinto de V^* el resultado del acuerdo no será igualitario. Sin embargo, conviene matizar esto. Pensemos por ejemplo en un caso en el que dos personas tienen la posibilidad de hacer un negocio cuyo beneficio bruto es, como en el ejemplo anterior, de 100 pesetas. Pero, a diferencia del caso anterior, para que el negocio salga adelante es necesario hacer una inversión inicial, pongamos por ejemplo, de 50 pesetas. Ninguno de ellos tiene mucho dinero, pero entre ambos consiguen juntar este capital inicial del modo siguiente: 1 pone 30

concesiones simultáneas, cada parte hará una concesión si y sólo si el hacerlo le ofrece una utilidad esperada mayor que la que tendría si se negara a ceder.

IV'. EFICIENCIA. Las dos partes estarán de acuerdo en hacer concesiones simultáneas si estas les proporcionan una utilidad esperada mayor que la que tendrían de otro modo.

Harsanyi muestra como a partir de estos postulados puede derivarse el principio de concesión de Zeuthen (demostración en Harsanyi, 36, pp. 150-1)

pesetas y 2, 20. Si el proceso de regateo se establece para llegar a un acuerdo sobre como repartir el beneficio bruto, el resultado no será igualitario. Concretamente, el resultado será (55,45). Este resultado es equivalente al que conseguiría del siguiente modo: primero se devuelve a cada uno de los jugadores su inversión inicial y posteriormente se establece el proceso de regateo para repartir lo que queda. Dicho de otro modo, el regateo tiene como objeto la partición del beneficio neto que se obtiene del negocio una vez que se han descontado los costes de cada una de las partes. Si se hiciera esto último, el resultado sería (25,25). Es decir, cada uno recibe en total lo que ha invertido más una parte igualitaria del beneficio neto, de modo que la diferencia entre lo que recibe 1 y lo que recibe 2 se debe a que cada uno ha invertido una cantidad distinta. En este ejemplo, entendemos que el negocio es imposible si no llegan a un acuerdo pero que, en caso de que este no se produzca, los dos jugadores no quedan en la misma situación. Esto se debe a que cada uno se queda con lo que tiene, y lo que tienen es distinto. Por eso puede interpretarse que el coste que tiene para 1 la estrategia de no cooperar es distinto, e inferior en este caso, que el coste que la no cooperación tiene para 2, si bien en términos absolutos ambos pierden lo mismo si no cooperan, a saber, 25 pesetas cada uno.

Este ejemplo tiene la finalidad de hacer notar que el carácter no igualitario del resultado de este tipo de casos es tan intuitivo como el resultado igualitario de los casos anteriores. Sin embargo, no todos los casos en los que U^* es distinta de V^* son tan claros. Esto se debe a que en muchos casos el resultado del juego, sea éste jugado o no de forma cooperativa, no consiste en pagos monetarios. Imaginemos un juego de dos personas en el que si, se desarrolla de un modo no cooperativo, 1 consigue una utilidad de $1/2$ y 2 de $3/8$, y en el que ambos pueden incrementar su utilidad si utilizan una estrategia conjunta. De entre todos los resultados óptimos posibles, y puesto que cada jugador debe conceder al otro al menos tanta utilidad como lograría de no cooperar, 1 propone el resultado $(5/8, 3/8)$ y 2 $(1/2, 1/2)$. Es decir, cada uno propone que el otro no gane nada con la cooperación mientras que él mismo se asigna el mayor incremento posible de utilidad. Si aplicamos el procedimiento N-2-H a este caso, el resultado será $(0.5625, 0.4375)$, que es claramente favorable a 1.

A pesar de las diferencias existentes entre este caso y el anterior, en ambos sucede básicamente lo mismo. En efecto, si en este último caso descontamos la utilidad inicial de las utilidades finales conseguidas, veremos que lo que resta es un valor igual para ambos, a saber, 0.0625. Si queremos emplear aquí los mismos términos económicos que se utilizaron en el ejemplo anterior, podemos decir que el beneficio neto de la cooperación se reparte igualitariamente. El problema es que en estos casos puede no resultar tan intuitivo el resultado. En el caso del negocio, parece claramente correcto que, antes de repartir nada, deben cubrirse los costos de ambos jugadores. Si el resultado del reparto del beneficio bruto no es igualitario, esto parece perfectamente natural: puesto que cada uno ha invertido una cosa, es natural que cada uno obtenga una cosa. En el otro caso la cuestión parece más discutible. Ninguno ha invertido nada. Lo único que ocurre es que cada uno obtiene una utilidad distinta del uso de estrategias no cooperativas. Sin embargo, las situaciones son estrictamente paralelas. Lo importante en ambos casos es la utilidad que cada uno de los jugadores obtendría en caso de no cooperación. Lo que garantiza el procedimiento N-Z-H en todos los casos es que el incremento de utilidad obtenido mediante la cooperación se distribuirá igualitariamente. Es decir, que cada jugador se beneficiará en la misma medida de la cooperación.

Cuando U^* es distinto de V^* lo que sucede es que las situaciones en las que se encuentran los jugadores no son simétricas. La diferencia entre ellos consiste en que cada uno obtiene una utilidad distinta en caso de no cooperación. Dicho de otro modo, uno pierde más que otro si no se llega a un acuerdo. Esto se traduce en la distinta fuerza con la que cada uno puede pretender un resultado favorable para él. El jugador cuya utilidad inicial es más alta puede conseguir un resultado de regateo favorable para él amenazando con no cooperar. El otro jugador sabe que ambos perderán si la amenaza se lleva a cabo, pero que él perderá más que el otro. Esto quiere decir que el jugador cuya utilidad es más alta en caso de no llegar a un acuerdo juega con ventaja en el proceso de regateo. El resultado no igualitario del regateo en estos casos refleja esta ventaja. Sin embargo, el procedimiento N-Z-H sigue siendo igualitario en un sentido. Este procedimiento, tal y como queda expresado en el principio de concesión, se interpreta habitualmente como una medida de las per-

didas relativas que cada jugador tendría de aceptar la propuesta del otro. Lo que el procedimiento garantiza es que las pérdidas relativas, o lo que es lo mismo, las concesiones relativas de cada uno de los jugadores sean iguales. Ahora bien, si la situación inicial de la que se parte no es simétrica, la concesión relativa sólo será igual si el resultado no es igualitario. Pero debe tenerse en cuenta que lo que no es igualitario es el punto de partida y no el procedimiento.

Como dijimos anteriormente (p.75) la solución al problema de regateo que supone el procedimiento N-Z-H puede aplicarse a la solución de juegos no estrictamente competitivos. Como se recordará, la idea básica consiste en inducir un juego de regateo G^* a partir de un juego no estrictamente competitivo G y tomar la solución al problema de regateo G^* como el acuerdo que permite una solución cooperativa a G . El juego de regateo G^* surge de colocar en el *statu quo* el par (U,V) que resultaría de G si este se desarrollara de forma no cooperativa.

Sin embargo, el problema surge precisamente en el momento en que hay que decidir qué es lo que debe tomarse como resultado del juego no cooperativo. La solución más simple es la conocida como "procedimiento de Shapley" y consiste en colocar en el *statu quo* de G^* el nivel de seguridad de las dos partes. Es decir, se supone que si el juego G se desarrollara de forma no cooperativa, cada uno de los jugadores utilizaría la estrategia pura o mixta que maximizará su nivel de seguridad (ver Cap.IV, p.60). En principio puede pensarse que este procedimiento es correcto y que en todos los casos da un resultado satisfactorio. En efecto, parece que colocar el nivel de seguridad de los jugadores en el origen es suficiente para asegurar que la fuerza de los jugadores quedara reflejada, ya sea esta fuerza igual o distinta. Es decir, parece que si el nivel de seguridad de ambos jugadores es el mismo esto indica que ambos ocupan posiciones simétricas y que las posiciones sólo serán asimétricas en el caso de que el nivel de seguridad de uno sea distinto al del otro.

Sin embargo, esto no es totalmente correcto. Hay casos en que las posiciones de los jugadores son asimétricas a pesar de que sus niveles de seguridad sean iguales. Un ejemplo de estas situaciones puede verse en Luce y Raiffa (57) p.139. Este ejemplo está representado por un juego con la siguiente matriz de pagos:

	B_1	B_2
A_1	(1,4)	(-1,-4)
A_2	(-4,-1)	(4,1)

Ejemplo 6

El nivel de seguridad de los jugadores en este juego es $(0,0)$. Si colocamos el par $(0,0)$ en el origen, el juego inducido de regateo G^* dará una solución igualitaria $(5/2, 5/2)$. Sin embargo, puede observarse en la matriz que, a pesar de que los niveles de seguridad de los jugadores son iguales, sus situaciones no son exactamente simétricas. Por el contrario, el jugador 2 tiene una ventaja sobre el jugador 1. Esta ventaja reside en el hecho de que 2 tiene una capacidad de amenaza de la que 1 carece. Supongamos que 2 amenaza con jugar la estrategia pura B_1 . Esto coloca a 1 en una situación difícil. Si 1 juega A_1 , entonces el resultado sería el mejor posible para 2. Por lo tanto, lo único que puede hacer 1 es amenazar a su vez con utilizar A_2 . Es decir, 2 puede forzar un resultado no cooperativo $(-4,-1)$ mediante la utilización de una amenaza conveniente. La amenaza de 2 es creíble porque, si bien el consigue un resultado malo si tiene que llevarla a cabo, hace que el contrario obtenga un resultado mucho peor, de hecho, el peor resultado posible para él. Es decir, 1 pierde mucho más que 2 si este cumple su amenaza. Por este motivo, puede resultar mucho más realista colocar $(-4,-1)$ en el origen del juego inducido de regateo G^* , pues 2 siempre podrá amenazar con este resultado si no se llega a un término satisfactorio de cooperación.

Nash establece un procedimiento que tiene la ventaja de ser sensible a las distintas capacidades de amenaza de los jugadores. Según este procedimiento, cada uno de los jugadores adopta una estrategia pura o mixta como amenaza. Puede darse el caso de que cada jugador sólo cuente con una estrategia de amenaza, o puede que estén en posición de elegir entre varias, en cuyo caso cada uno seleccionará una amenaza óptima, e.d., aquella que deje al contrario en la peor situación posible y que al mismo tiempo sea tan poco costosa como sea posible para uno mismo. El par de amenazas seleccionado por los jugadores establece un resultado (U,V)

que se utiliza como *statu quo* del juego de regateo, y se procede a continuación a resolver el problema de regateo de la manera indicada anteriormente.

Por lo que a nosotros respecta, lo que interesa es tener en cuenta lo que debe reflejar el *statu quo* del que parte el juego de regateo que servirá para lograr un compromiso cooperativo. Una función de solución satisfactoria que elija un punto de compromiso en cada juego no estrictamente competitivo que se presente debe tomar en cuenta la situación de la que los jugadores los jugadores parten, e.d., la situación en la que se encontrarían en caso de no cooperación, y que en la determinación de esta situación la capacidad de amenaza debe quedar reflejada*.

Hemos encontrado un procedimiento para lograr un acuerdo. En situaciones de juego de intereses mixtos, el resultado sólo será óptimo si los jugadores emplean una estrategia conjunta. Ahora bien, para esto deben ponerse de acuerdo en la estrategia a seguir, o, dicho de otro modo, en el resultado que producirá el uso de una estrategia conjunta. Si los jugadores son racionales, sólo estarán de acuerdo en un resultado tan bueno para ellos como sea posible, teniendo en cuenta que su beneficio esta limitado por una pretensión similar por parte del otro jugador. El procedimiento que hemos analizado selecciona un resultado, y por tanto una estrategia conjunta, para cada situación de este tipo.

En estas situaciones, lo racional es llegar a un acuerdo que tenga esta características. Y lo es porque el uso de esa estrategia conjunta les producirá un beneficio mayor que el uso de cualquier estrategia individual posible. Pero hasta ahora sólo se ha hablado de la racionalidad de llegar a un acuerdo. Queda aun por analizar la cuestión mucho más importante de si es racional cumplir el acuerdo. A este problema dedicaremos el capítulo siguiente.

* El procedimiento M-Z-H ha sido objeto de críticas y algunos autores han propuesto procedimientos alternativos. Este es el caso de Gauthier. Para una revisión de estas críticas y una discusión de este procedimiento alternativo, ver Apéndice 7.

CAPITULO VI

¿ES RACIONAL MANTENER LOS ACUERDOS?

En el capítulo anterior hemos analizado las situaciones en las que es racional llegar a acuerdos cooperativos, así como el tipo de acuerdos que aceptarían unos agentes racionales. Un agente racional sólo aceptará los acuerdos que le produzcan todo el beneficio que pueda conseguir, teniendo en cuenta que sus aspiraciones están limitadas por unas aspiraciones similares de los demás agentes. Partiendo de este principio, veíamos que un acuerdo aceptable por jugadores racionales debía consistir en una distribución equitativa de los beneficios de la cooperación. Esta condición se concreta en la exigencia de que el resultado del juego de regateo sea una partición igualitaria *siempre y cuando los jugadores desempeñen roles simétricos*. Cuando las situaciones en las que se encuentran los jugadores no son simétricas el acuerdo no será en este sentido igualitario, sino que reflejara la desigualdad del punto de partida. Dicho de otro modo, los beneficios de la cooperación se distribuyen igualitariamente en todos los casos, pero esto no conduce a una situación simétrica a menos que lo fuera la situación de partida.

Los acuerdos analizados en el capítulo anterior son acuerdos racionales. Ninguno de los jugadores podría conseguir algo mejor como fruto de la cooperación y son mejores para todos los jugadores que el resultado que se seguiría de una acción no cooperativa. Son acuerdos racionales pues son lo mejor que se puede hacer en una situación determinada en la que se encuentra el agente.

Una pregunta posterior a la de qué acuerdos son racionales es la de si es racional cumplir los acuerdos. Puede parecer que si un acuerdo es racional entonces es racional su cumplimiento. Sin embargo, esto no es así necesariamente. Podemos ver esto mediante un ejemplo.

En un país imaginario existe una sociedad esclavista, en la cual hay esclavos y hombres libres propietarios de esclavos. Los hombres libres viven bastante bien y los esclavos bastante mal. La vida en esta comunidad se desarrolla sin sobresaltos hasta que un joven de familia poderosa vuelve del extranjero donde ha cursado sus estudios y ha oído hablar de la teoría de la cooperación. A su vuelta, propone a sus mayores una alteración de su modo de vida. Su razonamiento es el siguiente.

El régimen esclavista bajo el que viven tiene muchos inconvenientes. A los hombres libres les resulta muy costoso el mantenimiento de la esclavitud y, por si esto fuera poco, reciben por parte de los esclavos unos servicios desgastados y poco esmerados. Por su parte los esclavos viven sin libertad bajo la presión del látigo y las cadenas. Tal situación es absurda y subóptima. Todos, amos y esclavos, podrían vivir mucho mejor. De modo que el joven propone hacer un trato con los esclavos. Los esclavos serán libres y recibirán un salario a cambio de sus servicios. Los amos pagarán estos servicios con poco más dinero del que les cuesta mantener el actual estado de cosas y recibirán servicios voluntarios y mucho mejores. Todos saldrán ganando, de modo que todos se comportarán de un modo racional si aceptan el trato. Sin embargo, los mayores no parecen muy convencidos por el razonamiento del joven. "Tu estas loco" le dicen. "Si hacemos lo que tu propones los esclavos no se convertirán en sirvientes voluntarios, sino que aprovecharán la menor oportunidad para romper el trato".

El ejemplo es originario de Gauthier¹ y plantea una situación en la que la racionalidad de un acuerdo y la de su mantenimiento no parecen coincidir. En efecto, el trato propuesto por el joven parece ser un trato razonable. En primer lugar, todo el mundo saldría ganando con el trato. Y, en segundo lugar, ninguna de las dos partes podría conseguir algo mejor. Los esclavos no accederían por menos de su libertad y un salario mínimo, y los amos no aceptarían nada peor que un servicio bueno, eficiente y sin malas caras. Para simplificar, podemos suponer que el trato es igualmente beneficioso para ambas partes, e.d., que todos ganan/conceden relativamente lo mismo. El acuerdo sería señalado como adecuado por los procedimientos analizados en el capítulo V. Por consiguiente, amos y esclavos se comportarían de un modo racional iniciando una negociación y llegando a este acuerdo. Sin embargo, los viejos del lugar tendrían razón al decir que no puede esperarse que los esclavos mantengan el trato. Es más, si los esclavos son agentes racionales no lo mantendrán. Con toda seguridad, una vez conseguida su libertad los ex-esclavos aprovecharán la menor oportunidad para forzar a los amos a un nuevo acuerdo, por ejemplo poniéndose en huelga.

¹ Gauthier (86, pp. 190-1)

El diagnóstico de Gauthier para este tipo de situaciones es el siguiente. Desde luego, el acuerdo es imparcial. Esto está garantizado por el método de regateo a través del cual se ha llegado al acuerdo. Pero sólo es imparcial respecto a la situación de la que se ha partido. El acuerdo conseguido a través del proceso de regateo es imparcial sólo en el sentido de que transmite sin alteraciones la imparcialidad del punto de partida. Pero del mismo modo transmite su parcialidad.

Parece que para saber cuando es racional mantener un acuerdo es necesario hacer una distinción entre lo que podríamos llamar acuerdos satisfactorios y acuerdos insatisfactorios. Esta distinción se establece según el punto de partida de la negociación, de tal modo que diremos que un acuerdo es satisfactorio cuando es el resultado de un proceso de regateo con un punto de partida satisfactorio e insatisfactorio en caso contrario. Esta definición no nos dice demasiado. Pero para poder decir algo más tenemos que tener antes un criterio para distinguir una situación de partida aceptable de otra que no lo es. A esta tarea dedicaremos la mayor parte de este capítulo.

I. Acuerdos satisfactorios

Sabemos que el resultado del proceso de regateo es un punto (U,V) cuya determinación depende del punto (U_0, V_0) que se tome como punto de partida. Y sabemos también que un acuerdo será adecuado si lo es el punto de partida. Habitualmente se considera que el punto de partida debe consistir en el resultado no cooperativo, e.d., el resultado que se seguiría si no se alcanzara un acuerdo satisfactorio para todas las partes. Sin embargo existían dos alternativas:

1- Por un lado esta la posibilidad de entender que, puesto que en el caso de que el juego se resuelva de un modo no cooperativo cada jugador utilizará la estrategia que maximice su nivel de seguridad, el punto de partida del regateo debe ser el nivel de seguridad de los jugadores.

2- Por otro lado, esta la alternativa planteada por Nash en su teoría de las amenazas óptimas.

Cuando planteamos ambas alternativas (pp.85) vimos que la propuesta de Nash parecía la más adecuada, pues reflejaba una asimetría en las posiciones de los jugadores que no siempre queda recogida en el resultado no cooperativo. De hecho, el resultado no cooperativo siempre puede ser entendido como una amenaza. En efecto, puesto que es el resultado que se producirá si no se llega a un acuerdo, los jugadores amenazan con jugar de modo no cooperativo si no se alcanza un resultado satisfactorio. Precisamente porque el resultado no cooperativo se utiliza como una amenaza el resultado del regateo no es igualitario cuando la situación en la que quedarían los jugadores en caso de no cooperación no es simétrica, ya que esto refleja la fuerza desigual con la que los jugadores plantean y defienden sus demandas.

La propuesta de considerar el resultado no cooperativo como punto de partida, ya sea entendiendo que este resultado será el nivel de seguridad de los jugadores o, alternativamente, que será la realización de las amenazas óptimas, esta sujeta a crítica. Esta crítica es doble. En primer lugar, se ha argumentado que la amenaza consistente en jugar de forma no cooperativa, en cualquiera de las dos versiones mencionadas y especialmente en la versión de Nash, no es creíble, puesto que de llevar-

se a cabo ambos jugadores saldrían perdiendo.

Sin embargo, esta crítica no parece muy adecuada. Por supuesto que todos salen perdiendo si se juega de modo no cooperativo. Precisamente este es el motivo por el que unos jugadores racionales están dispuestos a cooperar. Pero esto no significa que la amenaza de la no cooperación no sea creíble. Supongamos que dos jugadores quieren repartirse un millón de pesetas. Si cooperan, podrán repartirse esa cantidad. Si no cooperan, su lucha por conseguir lo más posible reducirá esta cantidad a la mitad (el resto iría, por ejemplo, a pagar a un abogado que defendiera sus pretensiones en un juicio legal). De este resto, el jugador 1 conseguiría 200.000 pesetas y el jugador 2, 300.000. Imaginemos que el jugador 1 propone un acuerdo según el cual él conseguiría 600.000 y su oponente 400.000. Si la amenaza del jugador 2 no fuera creíble, entonces este tendría que aceptar este resultado, pues con él gana más de lo que ganaría si llevara a cabo su amenaza de jugar de forma no cooperativa. Pero tampoco tendría más motivo para aceptar este resultado que el que tendría su oponente para aceptar un resultado de (300.000, 700.000), ya que, por el mismo motivo, la amenaza del jugador 1 tampoco sería creíble. De modo que si las amenazas de los jugadores no son creíbles, entonces ninguno tiene motivo alguno para aceptar un resultado en vez de otro. Podría pensarse que si los jugadores no pueden utilizar amenazas el resultado sería igualitario. Esto es dudoso. Supongamos que el jugador 1 propone (500.000, 500.000). El jugador 2 sólo aceptará este resultado en vez de otro más favorable para él si piensa que el jugador 1 no consentirá en un acuerdo peor para él. Pero ¿por qué no iba a consentir?, ¿por qué no iba a aceptar (300.000, 700.000)? De nuevo, el jugador 1 no puede forzar a 2 a aceptar otro resultado, y 300.000 es más de lo que él conseguiría sin cooperación. Parece por tanto que debemos admitir que el punto de partida del regateo debe ser considerado como una amenaza creíble².

A pesar de que el punto de partida deba considerarse siempre como

² Por supuesto, esto no significa que todas las amenazas son creíbles. Como dijimos en el capítulo anterior, una amenaza es creíble sólo si, a pesar de que ambos jugadores pierdan de llevarse a cabo, el autor de la amenaza pierde menos que el otro. Esto distingue una amenaza de un farol. En nuestro ejemplo, si no se llega a un acuerdo ambos salen perdiendo, pero el jugador 1 pierde más que el jugador 2. Por eso la amenaza de 2 es creíble y por eso el resultado de la cooperación será favorable a 2.

una amenaza, sigue en pie la cuestión de qué tipo de punto de partida hace que el resultado del regateo sea satisfactorio. En este sentido apunta la segunda crítica. Gauthier ha criticado la teoría tradicional sobre el punto inicial de regateo argumentando que los acuerdos basados en él no son satisfactorios y, por tanto, no son estables.

Esta crítica se basa en una confusión. Sin embargo, esta confusión no puede verse con claridad sin conocer la teoría de Gauthier acerca del criterio que debe utilizarse para distinguir un acuerdo satisfactorio de un acuerdo insatisfactorio. Por lo tanto, empezaremos por analizar esta teoría.

El acuerdo alcanzado mediante el proceso de regateo sólo será imparcial y racional si lo es la situación inicial de la que parte el regateo. Se trata por tanto de establecer las condiciones que hacen que la situación inicial de regateo sea imparcial y racional. Llamaremos a estas condiciones condiciones de negociación³.

La idea básica es que cada uno lleva a la mesa de negociación lo que ha conseguido o puede conseguir sin cooperación. Con una limitación: a nadie le está permitido mejorar su situación empeorando la situación de los demás. Con esta salvedad, cada agente tiene derecho a su propio cuerpo y sus capacidades, así como a los frutos conseguidos mediante la utilización de tales capacidades. Es decir, lo que cada uno puede llevar a la mesa de negociación es lo que puede conseguir en una situación en la que no exista interacción, lo cual significa que no sólo se debe considerar lo que se obtendría en ausencia de un acuerdo cooperativo, sino también que debe excluirse lo que se conseguiría "aprovechándose" de los demás.

Esta condición impuesta sobre la situación inicial de regateo tiene el efecto de evitar tanto el surgimiento de aprovechados⁴ como el de

³ Gauthier establece estas condiciones partiendo del análisis de la salvaguarda de Locke (1963) y realizando en este una serie de especificaciones y modificaciones.

⁴ Traduzco "free-rider" por "aprovechado" por creer que es el término castellano que más se aproxima a la idea original (nota 4). Las posibles alternativas, como "freeco" o "gorrón" tienen otras connotaciones no deseables, frente al término "aprovechado", que resulta mucho más neutro.

parásitos². Según argumenta Gauthier, una condición más fuerte que la salvaguarda, que exigiera que un agente debe mejorar la situación de los otros para poder mejorar la suya propia, facilitaría la aparición de aprovechados, del mismo modo que una condición más débil, que permitiera mejorar la propia situación a costa de un empeoramiento en la situación de los otros, daría lugar al parasitismo.

Ahora bien, podemos preguntarnos por qué la aceptación de la salvaguarda supone rechazar la teoría de las amenazas óptimas. El argumento fundamental contra esta teoría parece ser que las amenazas violan la salvaguarda por el hecho de ser amenazas. Desde luego, esto sucedería si las amenazas consistieran en amenazar a los demás con mejorar la propia situación a costa de empeorar la suya. Sin embargo, esto no tiene por qué ser así. Si la amenaza de un agente se lleva a cabo, sin duda esto supondrá un empeoramiento de la situación del otro jugador con relación a cual hubiera sido su situación en ausencia de interacción. Pero también supondrá un empeoramiento, aunque menor, de la propia situación. La amenaza no consiste en mejorar la propia situación a costa de empeorar la de el otro, sino en empeorar ambas.

Veamos un ejemplo. Imaginemos una isla en la que sólo hay dos habitantes, a los que llamaremos 1 y 2. 1 se dedica a la agricultura. Cultiva un huerto de hortalizas y un campo de árboles frutales. Por su parte 2 se dedica a la ganadería. Cría vacas, gallinas y corderos. Entre los dos habitantes no existe comercio ni ningún otro tipo de intercambio. Para vivir, 1 sólo cuenta con lo que cultiva, además de la caza eventual de animales no domesticados. De modo paralelo, 2 vive de los animales que cría y de la recolección de hierbas silvestres. Los dos habitantes tienen un conflicto desde hace tiempo. El conflicto surge debido a la escasez de pasto en la isla, lo que hace que el ganado de 2 necesite ocasionalmente invadir el huerto de 1. Un acuerdo cooperativo sería beneficioso

² La diferencia entre ambas figuras consiste en que mientras que el aprovechado obtiene un beneficio sin pagar los costos el parásito lo obtiene desplazando los costos a los demás. Utilizando los ejemplos de Gauthier (84, 87-90), aquellos que disfrutan de los beneficios de un faro erigido por otros son aprovechados, mientras que los dueños de fábricas que suelten gases en la atmósfera son parásitos. En decir, la diferencia está en la parte activa que el parásito desempeña tanto en la obtención del beneficio como en la generación de los costos. No obstante, la diferencia no resulta muy relevante ni fácil de establecer en la mayoría de los casos. El motivo de mencionarla aquí es simplemente señalar el por qué la condición sobre la situación inicial de regates no puede hacerse ni más débil ni más fuerte.

para ambas. La cooperación les permitiría intercambiar bienes, lo cual libraría a 1 de la caza, que se le da francamente mal, y a 2 de la recolección de hierbas, que además de ser una tarea desagradable para él, sólo le proporciona alimento vegetal de bajísima calidad. Además, podrían unir sus esfuerzos para construir un pozo con el que podrían regar la tierra y aumentar la zona fértil con lo cual producirían más productos vegetales para su propio consumo y podrían dedicar una zona a pasto para los animales. De modo que 1 y 2 viendo los beneficios de la cooperación, se sientan a negociar.

A la mesa de negociación cada uno viene con el resultado que obtiene en la situación no cooperativa. Ninguno ha violado la salvaguarda. Podría pensarse que 2 no lo ha respetado puesto que ha conducido su ganado a pastar en el huerto de 1. Esto no es así. Si 2 no hubiera hecho eso, no sólo no habría mejorado su situación, sino que la habría empeorado, pues sus animales habrían muerto. Además, si 1 no hubiera existido, 2 habría llevado de todas formas a sus animales a esa zona, que, como se recordará, es de las pocas zonas fértiles de la isla. De modo que 2 no se ha beneficiado del trabajo de 1 (a sus animales les es indiferente comer judías y tomates o comer la hierba que espontáneamente nace en esa zona).

Supongamos que viviendo de un modo no cooperativo ambos están en una situación paralela. Sin embargo 2 tiene una capacidad de amenaza de la que carece 1. En efecto, 2 puede amenazar con lo siguiente. Si no se llega a un acuerdo, 2 puede hacer que todo su ganado entre a la vez en el huerto y el campo de frutales de 1, arrasándolos. Esto les perjudicaría a ambos. 1 se quedaría sin más recurso que la caza, actividad peligrosa para la que 1 no tiene buenas condiciones. Es muy probable que 1 muriera. 2 también saldría perjudicado. Sus animales morirían y él se tendría que dedicar a la caza. Pero 2 es fuerte y rápido. Además sabe domesticar a los animales, y cuando el pasto volviera a crecer es posible que lograra alimentar a algunos animales. De modo que es muy probable que 2 sobreviva.

La cuestión es si 2 violaría la salvaguarda al llevar a cabo su amenaza. 2 no mejorará su situación si la amenaza se cumple. Esto ya es suficiente para mostrar que no se da ninguna violación de la salvaguarda. Pero, además, al realizar una amenaza lo que se está haciendo no es en llevar a la mesa de negociación algo que se ha conseguido o se puede con-

seguir en una situación no cooperativa, sino más bien mostrar los costos que cada uno podría hacer caer sobre el otro en caso de no cooperación. En cierto sentido, las amenazas desempeñan un papel puramente hipotético. Sin embargo, en otro sentido su papel no es hipotético. Si lo fuera, las amenazas perderían credibilidad, y por consiguiente, fuerza: tus amenazas pueden ser todo lo buenas que quieras, pero a menos que yo crea en la posibilidad de que se lleven a cabo no sirven para nada. Pero, en cualquier caso, el papel de las amenazas está esencialmente ligado a la negociación y por ello sólo cobran sentido una vez que el proceso de negociación se ha iniciado. Cuando la negociación se inicia las amenazas no se han llevado a cabo, y por tanto no pueden en ningún caso constituir una violación de la salvaguarda. En todo caso, podría plantearse si, en caso de que la negociación se rompiera y las amenazas se cumplieran, su cumplimiento violaría la salvaguarda. Pero es dudoso que pueda considerarse así. Y no sólo porque al cumplir la amenaza ambos resulten perjudicados, sino porque el simple hecho de romper la negociación ya perjudica a ambas partes, puesto que el resultado no cooperativo es peor para todos que la cooperación.

Podemos concluir que la utilización de amenazas no es incompatible con la aceptación de la salvaguarda. Es más, en este sentido, la utilización de amenazas no es incompatible con ninguna condición que se imponga sobre la conducta anterior a la negociación, pues las amenazas no son anteriores a la negociación.

Una vez aclarado esto, podemos pasar a analizar la teoría de Gauthier con el fin de ver si es aceptable. Para ello conviene que vemos más detenidamente porque se considera necesario imponer condiciones a la conducta previa a la negociación.

En ausencia de acuerdos, la conducta de un agente racional está dirigida con exclusividad a la maximización de su utilidad individual. Imponer condiciones a la conducta previa a la negociación significa restringir la libertad del agente para maximizar su utilidad. Por tanto, debemos preguntarnos por qué motivo un agente racional iba a aceptar una restricción de este tipo. Este motivo, a su vez, tendrá que estar relacionado con la maximización de la utilidad.

En una interacción no sujeta a restricciones y en la que cada agente trata de maximizar su utilidad surgen con facilidad la utilización

de la fuerza y el fraude. Los individuos más fuertes, más ágiles o más inteligentes utilizarán su superioridad para mejorar sus condiciones de vida a costa de los individuos menos afortunados. Presumiblemente, este es el origen de la sociedad esclavista del ejemplo. Pero, como vimos, si los amos intentaran llegar a un acuerdo con los esclavos partiendo de una situación inicial de regateo en la que se reflejara lo que cada uno tiene en el momento de iniciar la negociación, el acuerdo al que se llegaría sería tal que los esclavos no tendrían ningún motivo para mantenerlo. La nueva situación creada a partir del acuerdo sería tan desfavorable a los esclavos que estos no dudarían en romper el acuerdo tan pronto como fuera posible. Es más, puesto que los amos sabrían esto, sin duda nunca accederían a negociar con los esclavos*. Es decir, la negociación no sería posible, lo cual significaría que tanto amos como esclavos estarían peor de lo que podrían estar. Pero como todos están interesados en la negociación es presumible que todos estén dispuestos a partir de un punto que haga la negociación posible. Por consiguiente, los agentes racionales sólo aceptarán restricciones en tanto que estas posibilitan la realización de acuerdos beneficiosos. En otras palabras, sólo la perspectiva de un acuerdo beneficioso hace racional la aceptación de restricciones en la busca del propio beneficio previa al acuerdo, en tanto que tales restricciones son necesarias para posibilitar la negociación.

Aunque la aceptación de estas restricciones sólo tiene sentido en vistas de la negociación, puede plantearse si las restricciones deben aplicarse realmente o simplemente desempeñan un papel correctivo respecto a la situación que ha de ponerse como punto de partida del proceso de regateo. Llamaremos a estas alternativas A y B respectivamente. Esta cuestión puede plantearse:

- 1- respecto a la situación *anterior a la negociación*
- 2- respecto a la situación *posterior al proceso de negociación* y en caso de que este fracase.

* Esto no es del todo cierto. No se le escapará a nadie que como hecho histórico los amos han negociado con los esclavos en estos términos. También es un hecho histórico que los ex-esclavos han aprovechado todas las ocasiones para romper el acuerdo e intentar llegar a otro más ventajoso para ellos. Que lo hayan conseguido o no es discutible. La cuestión fundamental en este sentido parece ser la de si se tienen o no medios para forzar el cumplimiento de los acuerdos, así como del costo de estos medios. Más adelante nos ocuparemos de este tema.

1- A Esta alternativa parece ser la defendida por Gauthier y afirma que las condiciones deben respetarse siempre para que se pueda llegar a acuerdos satisfactorios. Esta alternativa tiene dos inconvenientes de considerable gravedad. En primer lugar, es un hecho que la interacción ha estado con frecuencia gobernada por la fuerza y el fraude. En estos casos, que me temo son la mayoría, ya se han violado las condiciones. Un ejemplo es el caso de la sociedad esclavista. Y ¿qué podemos hacer en estos casos? Parece que sólo hay dos caminos: 1) aceptamos la alternativa B para solucionar este tipo de casos, o 2) renunciamos a la posibilidad de que pueda haber en estas situaciones acuerdos satisfactorios.

Esta segunda posibilidad significaría que el establecimiento de la salvaguarda o de cualquier condición de este tipo sólo sirve como propuesta para casos futuros y no como criterio para juzgar los acuerdos pasados, que por definición se considerarían en su mayoría como insatisfactorios⁷.

Más grave y de peor solución es el segundo problema. Un agente racional sólo aceptaría unas condiciones que limitan su libertad para conseguir el mayor beneficio que le sea posible en tanto que estas condiciones son necesarias para posibilitar la aparición de acuerdos cooperativos satisfactorios. Pero esto limita en gran medida las situaciones en las que es previsible que se cumplan estas condiciones⁸.

⁷ Incluso su aplicación a casos futuros resulta dudosa bajo esta perspectiva. Supongamos que nos encontramos en la situación de nuestro ejemplo. Naturalmente, podemos decir que a partir de ahora no se volverán a violar determinadas condiciones. Pero esto no elimina los efectos de su violación previa. Por tanto, el futuro en el que podamos llegar a acuerdos satisfactorios es sin duda un futuro muy lejano. No tenemos posibilidad de llegar a acuerdos satisfactorios ni ahora ni a corto plazo y, muy probablemente, tampoco a un plazo medio. Esto de por sí resulta altamente insatisfactorio. De modo que parece que debemos admitir que para los casos actuales en los cuales ya ha habido una interacción bajo la presión de la fuerza y el fraude las condiciones que queramos imponer deben utilizarse como correctivos de la situación tal y como propone la primera alternativa. Sin duda, esto solucionaría el primer problema de esta alternativa, pero sólo a costa de renunciar a ella parcialmente.

⁸ Pongámonos en la situación de los años del ejemplo antes de convertirse en años. Forman una sociedad a la que llamaremos X. Ellos saben que si realizan una incursión en el poblado de los Y y los esclavizan (cosa que, dada su superioridad en materia bélica, podemos dar por segura) obtendrán una serie de beneficios. Puesto que los X son racionales, sólo renunciarán a estos beneficios si tiene la perspectiva de llegar con los Y a un acuerdo cooperativo cuyo resultado les sea aun más beneficioso. Que se produzca este acuerdo es algo de lo que no hay certeza. Por ello, en el cálculo deben entrar las probabilidades de que acontezca este suceso. Si las ventajas de la cooperación con los Y son muy superiores a las ventajas de esclavizarlos, entonces los X se conformarán con unas probabilidades relativamente pe-

Gauthier parece defender que la mera perspectiva de un acuerdo es suficiente para que un agente racional acepte determinadas restricciones en su búsqueda de beneficio. Esto no es defendible. Siempre existe alguna perspectiva, siempre es posible que se llegue a cooperar en algún momento. Pero esto no es suficiente. La perspectiva de la cooperación debe tener una probabilidad determinada. Esto no impide la utilización de la salvaguarda o de condiciones similares como algo que debe aplicarse con antelación a la negociación, pero si limita mucho los casos en los que tales condiciones serían racionalmente aceptables. En el resto de los casos estas condiciones se violarían. Si después se quiere llegar a un acuerdo, de nuevo habría que volver a admitir que en esos casos las condiciones actúan como correctivo para determinar una situación inicial de regateo satisfactoria.

Parece por tanto que los problemas de esta alternativa son lo suficientemente grandes como para que su aceptación sea problemática.

B. Puesto que se trata de un sociedad esclavista, sabemos que las restricciones no se han observado. Supongamos que los amos, queriendo llegar a una cooperación y sabiendo que esta no es posible si la negociación parte del actual estado de cosas, se disponen a hacer las modificaciones pertinentes. Para ello tienen que descontar de su haber lo obtenido mediante la violación de dichas restricciones*. Por su parte los es-

queñas de conseguir la cooperación. De todos modos, no hay mucho problema. Partiendo de una situación inicial de regateo adecuada y utilizando un proceso de regateo imparcial, si las ventajas son grandes para los X lo serán también para los Y, de modo que es muy probable que estos accedan a cooperar.

Más problemático es el caso en el que las ventajas de la cooperación sean pequeñas. Los X necesitarán una seguridad mayor para renunciar a tomar esclavos, e.d., necesitarán que la probabilidad de la cooperación sea mayor. Y puesto que las ventajas son menores, la probabilidad de cooperar no aumenta sino que disminuye. Pero todavía puede suceder algo peor. Puede suceder que la esclavitud les reporte más beneficios que la cooperación. Entonces, ¿tendrán los X algún motivo para renunciar al empleo de la fuerza y el fraude? Siempre es posible que en un futuro pueda surgir la posibilidad de cooperar. Pero seguramente esta posibilidad resulta e indeterminada no es suficiente para renunciar a unos beneficios seguros. Lo más probable es que sean sus hijos o sus nietos o sus bisnietos los que tengan la posibilidad de cooperar con los Y de un modo beneficioso.

* Más o menos su razonamiento podría haber sido el siguiente. Los amos se reúnen para realizar esta operación, y uno de ellos toma la palabra. "Si no hubiéramos hecho uso del fraude y la fuerza, sin duda no tendríamos esclavos. Esto significa que tendríamos que realizar el trabajo que ellos realizan ahora. ¿Podríamos hacerlo? De no ser así, tendríamos que descontar también los beneficios que obtenemos del trabajo de los esclavos, pues sin ellos no podríamos conseguirlos. Y aquí empezamos los problemas, porque yo, hablando sinceramente, sería incapaz de cortar leña, ir a por agua, trabajar en la plantación, realizar las tareas domésticas, cuidar del ganado, etc. Me tomo que varios siglos de inactividad física nos

clavos tendrían que descontar también los efectos de su esclavitud¹⁰. Todos, amos y esclavos, considerarán que la situación de partida es la que existía antes de que se produjera la interacción entre ambos grupos.

Esta utilización de las constricciones como correctivo de la situación de partida tiene también sus problemas. Posiblemente el fundamental sea el siguiente. Desde luego siempre es posible considerar que la situación de partida es aquella anterior a la violación de las condiciones que se quieren imponer. Pero esto no significa que esa situación sea la que existiría si tales condiciones nunca se hubieran violado. Tanto los amos como los esclavos del ejemplo habrían evolucionado de un modo distinto si nunca se hubiera dado una situación de esclavitud. Su situación podría ser mejor de lo que era cuando se violaron las condiciones, y también podría ser peor. Pero en cualquier caso no es posible saber que hubiera pasado. Las condiciones se han violado, y por mucho que se puedan descontar sus efectos no se puede volver atrás.

Sin embargo, este inconveniente sólo supone un obstáculo importante para la aceptación de esta alternativa si suponemos que la violación de esas condiciones ha impedido una mejora de la situación que existía anteriormente, *a.d.*, si suponemos que en el caso de que esas condiciones nunca se hubiesen violado la situación en la que originariamente se encontraban los distintos agentes habría mejorado. Afirmar esto no es lo mismo que afirmar que la violación de tales condiciones ha empeorado

han dejado un poco flojos. Esto sin contar con que yo no sé como se hacen todas esas cosas. Y es seguro que mi mujer no sabe cocinar tan bien como Anita. Y si la sugiero que tiene que ir al mercado y al río a lavar me enfrentará con una revuelta peor que la de los esclavos. No quiero ni pensarlo". En este momento interviene otro de los amos, cuya visión del asunto no es tan pesimista. "Todo esto está muy bien. Pero estás pasando algo por alto. Si no hubiéramos conseguido esclavos nosotros no nos hubiéramos apartado del trabajo físico ni nuestras mujeres de la cocina. De modo que podríamos hacer todo lo que ellos hacen ahora. De hecho, antes de que tuviéramos esclavos, nuestros antepasados cocinaban, cazaban, pescaban y cultivaban la tierra. Y no les debía ir tan mal, porque incluso consiguieron fuerza y recursos para fabricar las armas y los barcos con los que conseguimos los esclavos. Todo esto sin contar el arte, la literatura, la filosofía y la ciencia política, que, si bien ahora florecen más debido al tiempo que nos deja libre el trabajo de los esclavos, antes tampoco eran despreciables. De modo que yo propongo que sea esto lo que se considere como punto de partida. Si hay que descontar los efectos de la fuerza y el fraude hay que descontarlos todos, es decir, no sólo los beneficios que eso nos ha producido sino también lo que nos ha quitado en lo que ha fuerza física y conocimientos prácticos su reflejo".

¹⁰ La situación de la que partirán será una situación en la que ellos son libres. Pero también es una situación en la que ellos carecen de los conocimientos y la fuerza física que les ha dado la esclavitud.

la situación de alguien mejorando la de otros. Esto último es lo que ha tenido que suceder para que pueda hablarse de la violación de unas condiciones necesarias para que puedan darse acuerdos satisfactorios. Podemos decir que, de haberse respetado las condiciones, la situación de algunas personas sería mejor de lo que es ahora. Pero esto no significa que pueda afirmarse que, de no haberse cometido tal violación, la situación en la que ahora se encontrarían algunas personas sería mejor de la que era en el momento en que se incumplieron las condiciones¹¹.

No tiene sentido hablar de lo que habría pasado si no hubiera pasado lo que ha pasado. No sólo no tiene sentido, sino que tampoco tiene importancia. Lo realmente importante es que, para que el acuerdo resulte satisfactorio, tu no puedes darme mediante el trato lo que me has quitado antes. Esto significa que, si queremos hacer un trato satisfactorio, primero me tienes que devolver lo que me has quitado, pero no que me tengas que dar lo que yo tendría si nunca me hubieras quitado nada.

Podemos concluir que la alternativa B expresa la forma más adecuada de interpretar las condiciones que deben imponerse para que sea posible la realización de acuerdos satisfactorios. Por tanto, diremos que la situación que se toma como punto de partida en el proceso de regateo debe ser una situación que no viole esas condiciones, con independencia de que la situación de hecho que existe en el momento de iniciar la negociación suponga o no una violación de las mismas.

2- La segunda cuestión trata sobre qué sucederá si mediante la negociación no se consigue un acuerdo cooperativo. Naturalmente, en este caso la interacción de desarrollará de un modo no cooperativo. Pero queda por aclarar si en esta interacción no cooperativa se respetarán o no las condiciones establecidas.

Supongamos que en nuestro ejemplo amos y esclavos inician un pro-

¹¹ Pensemos en los esclavos del ejemplo. Supongamos que la situación actual comenzó hace 100 años. Si comparamos su situación actual con su situación anterior a la violación de la salvaguarda veremos que ha empeorado. Paralelamente, la situación de los amos es ahora mejor de lo que era antes. Pero esto no significa que, si la salvaguarda siempre se hubiera respetado, los esclavos estarían ahora mejor de lo que estaban hace 100 años, del mismo modo que tampoco significa que los amos estarían peor. Podría haber sido así, pero ¡quien sabe!. Los amos podrían haber descubierto petróleo durante esos 100 años y los esclavos podrían haber sufrido una epidemia mortal que los hubiera diezmado.

ceso de regateo en el que se parte de una situación en la cual se han descontado los efectos de la violación de dichas condiciones. Y supongamos que por el motivo que sea amos y esclavos no consiguen llegar a un acuerdo. Podemos preguntar que sucedería en ese caso. En principio se dan dos posibilidades: A amos y esclavos se quedan en la situación que se ha tomado como punto de partida de la negociación o B se quedan como estaban¹².

A Puesto que esta situación no coincide necesariamente con la situación real, lo primero que debemos analizar es si es posible que tal situación se realice, e.d., si es posible que la situación hipotética se haga real. La respuesta a esta primera cuestión es negativa. No siempre es posible realizar la situación hipotética¹³. Naturalmente, esto no tiene por que suceder en todos los casos¹⁴. La diferencia entre ambos tipos

¹² Nótese que si hubiésemos escogido la alternativa A como interpretación correcta de las condiciones de negociación este problema no se plantearía. Puesto que en ese caso la negociación sólo sería posible si nunca se hubiese dado una violación de esas condiciones, en caso de que no se alcanzara un acuerdo cooperativo la situación que respecta las condiciones y la situación realmente existente antes del inicio de la negociación coincidirían. Sin embargo, una vez que hemos aceptado la alternativa B esa coincidencia, aunque puede darse, no se da necesariamente. Porque, según esta interpretación de las condiciones de negociación, puesto que es posible que no se hayan respetado las condiciones en la interacción anterior a la negociación, la situación que se toma como punto de partida del regateo es en la mayoría de los casos una situación hipotética que no coincide con la situación de hecho.

¹³ Recordemos lo dicho en el caso de la sociedad esclavista (ver notas 6 y 7). Descontar los efectos de la violación de las condiciones de negociación no sólo supondría devolver la libertad a los esclavos, cosa sin duda posible en todo momento, sino también anular todos los efectos que sobre amos y esclavos han tenido cien años de un modo determinado de vida. Y esto, desde luego, no es tan fácil. Los amos no pueden recuperar con un sólo acto de voluntad sus conocimientos prácticos ni su fuerza física, ni sus mujeres sus cualidades culinarias. Los esclavos tampoco pueden olvidar lo aprendido ni recordar lo olvidado. Nada puede volver a ser lo que era.

¹⁴ Imaginemos una situación ligeramente distinta a la del ejemplo. En esta nueva situación nos encontramos también con una sociedad esclavista en la cual se han violado las condiciones de negociación, pero que, al contrario de lo que sucedía en el caso anterior, es una situación reciente. Los X acaban de capturar a los Y, de modo que podemos decir que sólo nominalmente son amos los primeros y esclavos los segundos. Digo nominalmente porque en realidad tanto unos como otros son los mismos que eran antes de la violación de los condiciones de negociación. Para hacer la cuestión más sencilla, supongamos además que la batalla ha sido incruenta y sólo se han producido daños materiales fácilmente reparables. Después de la captura, los amos se reúnen en su campamento para considerar la situación. Haciendo cálculos, los amos se dan cuenta de que la situación que se producirá a partir de ahora será menos que óptima, y que, por tanto, será mucho mejor para todos negociar. Es fácil comprobar que en este caso no surgirán los problemas que vimos en el caso anterior. La situación hipotética que debe tomarse como punto de partida de la negociación, e.d., la situación anterior a la violación de las condiciones de negociación, es la situación de ayer y, de ayer a hoy, no ha sucedido nada irreparable. Todo lo que hay que hacer es liberrar a los esclavos y pagar los costos de unas cuantas cabañas. Borrón y cuenta nueva.

de casos reside en cuestiones de hecho, tales como el tiempo transcurrido desde la violación de las condiciones o lo amplios y profundos que hayan sido sus efectos. Aparte de estas generalidades, no hay ninguna manera de decidir en general cuando será posible realizar la situación hipotética y cuando no. Se trata de una cuestión que sólo puede decidirse a la vista del caso concreto.

Que no siempre sea posible recuperar la situación hipotética ya es de por sí un inconveniente con el que debe enfrentarse cualquiera que quiera defender esta postura. Sin embargo, este inconveniente no es necesariamente decisivo. Podría pensarse que en los casos en los que es posible debe hacerse y, en los casos en los que no es posible, al menos debe hacerse en esa dirección tanto como se pueda¹².

Es importante señalar que esta conducta a la que todos se comprometen al iniciar la negociación no forma parte del acuerdo cooperativo al que se pretende llegar. De hecho, ni siquiera consiste en la realización de un resultado óptimo. Más bien, esta conducta, o mejor dicho, el compromiso de seguir esa conducta en caso de que no haya acuerdo, debe tener la misma consideración que tienen las condiciones de negociación, de las cuales no son sino su realización, a saber, deben considerarse como condiciones de posibilidad de una negociación que tenga como resultado un acuerdo satisfactorio.

Desde este punto de vista, la postura que estamos considerando puede defenderse del siguiente modo. Queremos negociar para conseguir un acuerdo óptimo. Pero además queremos que el acuerdo sea satisfactorio, y para que esto sea posible son necesarias dos cosas. En primer lugar, es necesario que el proceso de regateo tenga como situación de partida una

Esto significa que en este caso, si la negociación fracasa, es posible plantear la posibilidad de que la situación no cooperativa sea la indicada en el punto de partida del regateo. Todo el mundo se queda como estaba antes de que los X capturaran a los Y, y, como suele decirse, aquí no ha pasado nada. Pero en otros muchos casos, tal y como sucede en el ejemplo de la sociedad esclavista con solera, esto sencillamente no es posible.

¹² Por ejemplo, en el caso de la sociedad esclavista consolidada podría hacerse algo parecido a esto. En el caso de que no se llegue a un acuerdo cooperativo, las partes involucradas en la negociación harán lo siguiente. Los años devolverán la libertad a los esclavos y les pagarán un viaje de vuelta a su lugar de origen. Los esclavos "devolverán" a los años los conocimientos que estos les dieron. Los enseñarán a llevar la casa y las tierras etc. Más no puede hacerse, pero quizá esto sea suficiente.

situación en la que no se tomen en cuenta las ventajas obtenidas por las partes como resultado del empleo de la fuerza y el fraude. En *segundo* lugar, es necesario garantizar que, en caso de no llegar a un acuerdo satisfactorio, la situación en la que quedarán las partes será, en la medida de lo posible, no la situación real, sino la situación indicada en el punto de partida de la negociación, porque sólo comprometiéndose con la realización de esta situación las partes estarán realmente libres de la presión de la fuerza y el fraude.

Debe notarse que estos dos supuestos son lógicamente independientes y que, por tanto, la justificación de uno no implica la justificación del otro. Hasta ahora todo lo que hemos visto es que, para que sea posible un acuerdo satisfactorio, debe satisfacerse el primero de estos supuestos. Pero esto sucede con independencia de lo que pueda suceder en el caso de no conseguir llegar a un acuerdo. Veamos esto en el ejemplo de los amos y los esclavos.

Supongamos que ambos grupos se sientan a negociar. Puesto que quieren llegar a un acuerdo satisfactorio, se disponen a establecer un punto de partida una situación en la que los efectos de la violación de las condiciones de negociación no aparezcan. Supongamos que la utilidad de esa situación hipotética es de 30 para los esclavos y de 50 para los amos. Si llegan a un acuerdo, podrán producir un bienestar superior, digamos de una utilidad de 150. Empleando el procedimiento de regateo analizado en el capítulo anterior, se llegaría a un acuerdo sobre un resultado de (65, 85), donde 65 sería la utilidad de los esclavos y 85 la de los amos. Este acuerdo es satisfactorio puesto que por hipótesis hemos partido de una situación inicial de regateo en la que se respetan las condiciones de negociación y hemos desarrollado la negociación según un procedimiento de regateo imparcial. Supongamos que la situación real en la que se encuentran amos y esclavos tiene una utilidad de 60 para los primeros y de 20 para los segundos. Desde luego, si hubiéramos admitido esta situación como punto de partida de la negociación, el resultado del regateo habría sido distinto. Además, por hipótesis no hubiera sido satisfactorio. Pero, puesto que no lo hemos hecho, la situación real no ha tenido ninguna influencia en la determinación del resultado.

Precisamente, esto es lo que niegan los defensores de la postura que estamos analizando, *e.d.*, los que mantienen que en caso de no llegar

a un acuerdo debe garantizarse que el resultado no cooperativo será, no la situación real, sino la situación hipotética utilizada como punto de partida de la negociación (es decir, los que mantienen que, en nuestro ejemplo, la no cooperación tendrá un resultado de (30,50) y no de (20,60)). Y lo niegan alegando que esto es necesario para llegar a un acuerdo satisfactorio. Pero afirmar esto significa afirmar que el resultado real no cooperativo influye en la determinación del resultado de la negociación a pesar de que no desempeña ningún papel en el proceso de regateo (¿Cómo es esto posible? Puede suceder que se este pensando en algo similar a esto. Si todos sabemos que, en caso de no alcanzar un acuerdo, el resultado será (20,60) y no (30,50) entonces los esclavos tenderán a conceder más de lo necesario porque los amos, por su parte, sabrán que pueden apretar más de lo que teóricamente les está permitido. Es decir, aunque en teoría se regatee con un punto de partida de (30,50), la situación que realmente está operando en el fondo es la situación real de (20, 60). Esta será la que influya decisivamente en el resultado, y como en ella no se respetan las condiciones de negociación, el acuerdo no será satisfactorio.

Tomando como punto de partida el resultado (30,50) el regateo desembocará en un resultado de (65,85). Este resultado será el admitido ya que, supuesta la racionalidad de las partes, ninguna puede esperar conseguir más ni tampoco se conformará con menos. Ahora bien, adoptemos por un momento el punto de vista de los amos. Aunque ellos no puedan conseguir más de 85 en esta negociación, podrían conseguir más si el punto de partida fuera (20, 60). Y puesto que si no se llega a un acuerdo ese será realmente el resultado, la presión que realmente pueden ejercer obedece a su posición en este resultado y no en el resultado hipotético de (30,60). Luego parece necesario establecer un compromiso para que, en caso de no cooperación, el resultado sea (30,60), ya que sólo así puede asegurarse que esta sea el único resultado que influya en la negociación.

Este razonamiento destinado a mostrar la necesidad de hacer coincidir el punto de partida de la negociación con el resultado real no cooperativo pasa por alto algunos factores relevantes. En primer lugar, no tiene en cuenta que, una vez admitido un punto de partida de la negociación, el proceso de regateo señala inequívocamente un resultado como aquel en el que los agentes racionales consentirán. Los amos no intenta-

rán conseguir más porque eso sería pretender que los esclavos hagan una concesión relativa mayor que la suya, cosa que nunca harán supuesto que son agentes racionales. Los amos sólo podrían forzar a los esclavos a aceptar un resultado distinto realizando el cálculo a partir de un punto distinto. Pero, de hacer eso, la negociación sería otra distinta, con otro punto de partida y, por consiguiente, con otro resultado. Es decir, no hay ninguna manera en la que un resultado pueda influir en el proceso de regateo bajo cuerda. Una vez establecidos los parámetros que definen un juego de regateo, el procedimiento de resolución es tal que no admite nada "psicológico" en su desarrollo¹⁴.

Si de todos modos queremos defender esta postura aun podemos decir otra cosa a su favor. Podemos decir que lo que sucedería es que los amos de nuestro ejemplo no admitirían como punto de partida de la negociación la situación hipotética, sino que forzarían a los esclavos a negociar en base a la situación real no cooperativa. Y aquí es donde aparece el *segundo* factor que el razonamiento a favor de esta postura pasa por alto. Naturalmente que los amos *pueden* forzar a los esclavos a negociar sobre el resultado real obtenido mediante la utilización de la fuerza y el fraude, y pueden hacerlo porque pueden amenazar con esta situación, que es la que *realmente* se producirá si no se llega a un acuerdo cooperativo. Pero los amos saben que, si hacen esto, nunca llegarán a un acuerdo satisfactorio. Y puesto que estamos suponiendo que es un acuerdo de este tipo lo que quieren conseguir, nunca utilizarán su capacidad de amenaza para establecer un proceso de regateo con un punto de partida insatisfactorio. Por el contrario, estarán dispuestos a negociar a partir de una situación en la que los efectos de la violación de las condiciones de negociación haya sido anulado. Por hipótesis esto ya ha sucedido. Pero entonces ya han renunciado a negociar sobre la base de la situación real. Pueden volverse atrás y querer cambiar la base de la negociación. Pero entonces estarían renunciando a un acuerdo satisfactorio.

Hasta ahora hemos visto que la alternativa A no es muy realista.

¹⁴ Esto se vera con más claridad si pensamos que, tras determinar los datos relevantes, las partes pueden irse a su casa o a charlar en el bar mientras que una máquina resuelve el problema en su lugar. Lo único que tienen que hacer es ir al cabo de un rato a recoger el resultado. De modo que si alguna de las partes cree que puede forzar el resultado más a su favor lo único que puede hacer es cambiar el punto de partida de la negociación o algún otro dato que pueda influir en este sentido.

es innecesaria y poco eficaz. Nos queda por analizar si sería racional admitir tal exigencia.

Por definición, en una situación en la que se han violado las condiciones de negociación hay un agente o un grupo de agentes que se han beneficiado a costa de otros. Es decir, hay alguien para quien la situación es mejor si se violan esas condiciones que si no se violan¹⁷. De modo que lo que debemos preguntar es si los que se benefician de esta violación obrarían racionalmente al admitir esta exigencia.

Un agente racional sólo renunciará a los beneficios así obtenidos si el hacerlo resulta ser una condición necesaria para posibilitar la realización de un acuerdo cooperativo más beneficioso aun. Para lograr un acuerdo satisfactorio es necesario que se renuncie a utilizar esos beneficios como punto de partida de la negociación. Pero una vez que se admite esto, la exigencia de que en caso de no cooperación esa situación hipotética se convierta en la situación real no tiene ninguna influencia en el resultado del regateo. Por tanto, si admitieran esa exigencia, lo único que harían sería empeorar su situación en caso de no cooperación sin que esto tuviera ninguna contrapartida.

Podría argumentarse que, aun en el caso de que no se lograra llegar a un acuerdo, el respeto de las condiciones de negociación en la situación no cooperativa posterior tendría el efecto de posibilitar futuros acuerdos, efecto que sería sin duda beneficioso. Este argumento retrotrae la discusión a un punto ya discutido, a saber, si es necesario que las condiciones de negociación se respeten con anterioridad a la negociación. Puesto que ya vimos los argumentos en contra de esta postura (pp.97 y ss.) no creo necesario volver a repetirlos.

B. Puesto que hemos rechazado la alternativa A, tenemos que aceptar que lo importante para que sea posible un acuerdo cooperativo satisfactorio no es comprometerse a respetar las condiciones de negociación si el acuerdo no se logra. Las condiciones de negociación son justamente eso, condiciones establecidas con vista a la negociación y que tienen la

¹⁷ Naturalmente, se entiende que se habla de la situación en ausencia de acuerdos cooperativos. Un agente racional se comprometerá con este tipo de acuerdos si el hacerlo resulta más beneficioso que cualquier otro posible curso de acción no cooperativa. De modo que el dilema de la renuncia a los beneficios obtenidos mediante la violación de las condiciones de negociación sólo se plantea cuando la cooperación es más beneficiosa que una situación no cooperativa en la que se violen dichas condiciones.

finalidad de conseguir que esta tenga como resultado acuerdos satisfactorios. Y del mismo modo que vimos que no es necesario que esas condiciones se respeten antes de la negociación, tampoco es necesario que se respeten después si la negociación fracasa, sin que esto suponga que el resultado de la negociación este influido por la situación real no cooperativa en la se violan estas condiciones¹⁴.

Podemos concluir que la postura más adecuada respecto a las condiciones de negociación es considerarlas como algo que, no sólo no adquiere sentido más que con la perspectiva de un acuerdo, sino cuya función esta restringida a actuar como correctivo de la situación no cooperativa real a fin de conseguir un punto de partida de la negociación que permita alcanzar acuerdos satisfactorios. Y esto es así, básicamente, porque que un acuerdo sea o no satisfactorio depende exclusivamente del carácter intrínseco del acuerdo, que a su vez esta determinado únicamente por la situación de partida del proceso de regateo y por el conjunto de resultados óptimos entre los que los agentes tienen que escoger.

Puesto que ya sabemos cómo deben interpretarse las condiciones de negociación y que papel deben desempeñar, podemos pasar a analizar por extenso por qué, y en casos, un agente racional debe aceptar tales condiciones.

¹⁴ Puede pensarse que, de todos modos, los esclavos estarían más tranquilos si los amos se comprometieran a respetar las condiciones de negociación fuera cual fuera el resultado. Sin embargo esto no tiene ningún fundamento. Los amos, al aceptar un punto de partida satisfactorio y al sentarse a negociar ya se comprometen a aceptar el resultado de la negociación, siempre y cuando este sea imparcial y racionalmente aceptable, cosa que podemos dar por segura dado el procedimiento de regateo que manejan. Un compromiso adicional no añade ninguna seguridad. Al fin y al cabo, si los amos no respetan su compromiso de aceptar el resultado de la negociación que parte de una situación satisfactoria, ¿porqué iban a respetar el compromiso de realizar esa situación en todo caso? No sólo la garantía no es mayor, sino que es mucho menor. El acuerdo cooperativo beneficia a los amos al igual que a los esclavos, mientras que el respeto de las condiciones de negociación en ausencia de un acuerdo les perjudica. De modo que si no respetan el acuerdo cooperativo, que les resulta beneficioso difícilmente puede esperarse que respeten un compromiso del cual no extraen ningún beneficio.

II. Por qué aceptar las condiciones de negociación

En el apartado anterior dijimos que el propósito de las condiciones de negociación era posibilitar la creación de acuerdos satisfactorios. En su momento, caracterizamos un acuerdo satisfactorio como aquel que surge de un proceso de regateo cuyo punto de partida es satisfactorio. A su vez, un punto de partida será satisfactorio si respeta las condiciones de negociación. Sin embargo, esta caracterización se limita a conectar unos conceptos con otros sin explicar ninguno de ellos. Por tanto, es preciso definir de modo independiente lo que es un acuerdo satisfactorio.

Al comienzo de este capítulo establecimos una distinción entre acuerdos racionalmente aceptables y acuerdos que, además, un agente racional desearía mantener. Es a estos últimos a los que nos referimos con el nombre de "acuerdos satisfactorios". Por ello, preguntar qué condiciones de negociación deben imponerse es preguntar en qué condiciones estaría un agente racional dispuesto, no sólo a aceptar un acuerdo, sino también a desear su cumplimiento.

No sólo sabemos el propósito que las condiciones de negociación deben cumplir. También sabemos algo acerca de la característica general que han de tener estas condiciones. Esta característica deriva directamente de la definición de racionalidad práctica. Un agente racional busca la maximización de su utilidad y sólo acepta restricciones a su conducta en tanto que esto es un medio necesario para lograr esa maximización. Un agente racional aceptará determinadas restricciones supuesto que estas son necesarias para la creación de acuerdos satisfactorios, pero sólo en esta medida. Por ello, las condiciones de negociación deberán ser las restricciones mínimas suficientes para conseguir ese propósito.

Gauthier ofrece una teoría completa acerca de cuales deben ser esas condiciones de negociación (86.cap.VII)¹⁹. Tal teoría puede resumirse diciendo que el punto de partida de la negociación debe respetar el

¹⁹ Al comienzo de este capítulo (pp.93 y ss.) apuntamos algunos rasgos generales de esta teoría. Para un análisis extenso de la propuesta de Gauthier respecto a las condiciones de negociación, ver Apéndice B.

principio de que cada uno de los agentes tiene derecho a todo lo que pueda conseguir con su propio trabajo siempre y cuando con ello no empeore la situación de los demás, a menos que esto sea necesario para evitar empeorar la situación propia. (En adelante, nos referiremos a este principio con el nombre de "salvaguarda lockeana"²⁰) La salvaguarda lockeana constituye para Gauthier la condición de negociación mínima suficiente que posibilita la creación de acuerdos satisfactorios. Es decir, ningún agente racional consideraría satisfactorio un acuerdo cuyo punto de partida violara las condiciones de la salvaguarda. La defensa que Gauthier hace de la salvaguarda se basa principalmente en mostrar que su aceptación elimina tanto los parásitos como los aprovechados, lo cual parece ser condición necesaria para garantizar la imparcialidad de los acuerdos.

La argumentación en este sentido consiste en considerar lo que ocurriría si exigiéramos unas condiciones de negociación o bien más fuertes que la salvaguarda o más débiles. Una condición más fuerte establecería que un agente sólo puede mejorar su situación si con ello mejora también la de los demás²¹. El cumplimiento de esta condición supondría que determinadas personas verían mejorada su situación a costa del trabajo o las capacidades de otras. Es decir, no sólo facilitaría la aparición de aprovechados sino que la exigiría. Por otro lado, una condición más débil permitiría a un agente mejorar su situación a costa de un empeoramiento de la situación de los demás. La situación de los demás empeoraría ya que alguien se aprovecharía de su trabajo y sus capacidades, o lo que es lo mismo, porque alguien estaría actuando como un parásito²².

²⁰ El origen y razón de ser de este nombre queda explicado en el Apéndice 8.

²¹ Una condición de este tipo es la que establece Rawls en su segundo principio de justicia. No es este el momento de discutir si el principio de Rawls es preferible al de Gauthier, pues se trata en principio de dos condiciones de orden distinto. La condición de Gauthier es una condición de negociación en el sentido explicado, mientras que el principio de Rawls es un principio de carácter moral. No obstante, conviene tener presente la diferencia entre ambas condiciones para una discusión posterior.

²² Que una condición más débil o más fuerte que la salvaguarda en el sentido indicado daría lugar a la aparición de parásitos y aprovechados es difícilmente discutible. Pero para que esto suponga una defensa de la salvaguarda es necesario mostrar que ambos fenómenos son, desde el punto de vista de un agente racional, indeseables. Para Gauthier la indeseabilidad de estos fenómenos surge del hecho de que en ambos casos la imparcialidad de la situación queda afectada. En el caso del parásito, esto sucede porque la situación de los demás queda directamente empeorada. En el caso del aprovechado porque, si bien su existencia no empeora directamente la situación de los demás, sí lo hace de modo indirecto, ya que él obtiene un beneficio cuyos costos han pagado otros. Veamos dos ejemplos. Supongamos que un grupo

Una situación en la que existen parásitos o aprovechados no es imparcial. Y no lo es porque en ambos casos algunas personas se benefician a costa de otras. Sin embargo, que la salvaguarda impida la aparición de estos fenómenos sólo aporta en su defensa un apoyo indirecto, en tanto que hace plausible la pretensión de la salvaguarda de constituir una condición de negociación razonable. Pero la plausibilidad no es una prueba. Aun queda por demostrar que un agente racional estaría dispuesto a aceptar las condiciones de la salvaguarda y que no aceptaría como satisfactorio un acuerdo basado en unas condiciones distintas.

Conviene hacer notar que la posición de un parásito o un aprovechado es, desde el punto de vista de un agente racional, envidiable. Ocupar uno de estos lugares significa conseguir determinados beneficios de manera gratuita. Desde luego, desde el punto de vista de la maximización de la utilidad, esto es un *desideratum*. Aceptar unas condiciones de negociación que hagan difícil que alguien ocupe estas posiciones significa renunciar a estar uno mismo en tan envidiable situación. Pero, si esta situación es tan deseable, ¿por qué un agente racional va a renunciar a ella?

Para contestar esta pregunta, volvamos de nuevo al caso de los amos y los esclavos de nuestro ejemplo. Establecer un regateo con un punto de partida que respete lo obtenido mediante la violación de la salvaguarda les favorece enormemente. Sólo aceptarán modificar el punto de partida de la negociación según las exigencias de la salvaguarda como un fin necesario para lograr un acuerdo satisfactorio.

Aceptar las exigencias de la salvaguarda es necesario para conseguir un acuerdo satisfactorio *desde el punto de vista de los esclavos*. Un acuerdo satisfactorio es un acuerdo que un agente racional está dispuesto

de personas, al que llamaremos X, trabaja cultivando la tierra. Otro grupo, el de los Y, propietarios de fábricas, arrojan sustancias contaminantes sobre los campos de los X. La situación de los X empeora directamente por las acciones de los Y. Los Y son parásitos. Veamos ahora una situación en la que se da el otro fenómeno. Supongamos que esta vez los X construyen carreteras. Para este fin, trabajan e invierten. Los Y no hacen nada de esto, pero se benefician de la existencia de las carreteras igual que los X. Viajan por ellas, transportan sus mercancías y reciben los beneficios que otro grupo, los Z, dejan en sus viajes turísticos al país de los X y los Y, viajes que emprenden anhelados por la comodidad de sus carreteras. En este caso, la situación de los X no empeora en términos absolutos por el comportamiento de los Y. Pero sí que empeora comparativamente. Para obtener los beneficios que dan las carreteras, los X pagan determinados costos. Los Y obtienen los mismos beneficios sin ningún costo por su parte. Los Y son aprovechados.

a cumplir. Y los amos estarían dispuestos a cumplir un acuerdo basado en una violación de la salvaguarda. Naturalmente, esto sucede porque ellos son los beneficiados por ese incumplimiento. Por ese mismo motivo, los esclavos no encontrarían satisfactorio ese acuerdo, y por ello, estarían dispuestos a romper el acuerdo a la menor oportunidad. La cuestión es saber porque los amos deben estar interesados en que el acuerdo resulte satisfactorio para los esclavos. No debe olvidarse que partimos del supuesto de que los amos no tienen ningún interés en el bienestar de los esclavos²³. El único motivo que puede llevar a los amos a desear que el acuerdo sea satisfactorio desde el punto de vista de los esclavos es que desean un acuerdo que estos estén dispuestos a mantener. A su vez, la único razón que puede justificar este deseo es que en caso contrario, los amos, para garantizar su cumplimiento tendrían que emplear unos métodos coercitivos cuyos costos de mantenimiento sobrepasarían los beneficios obtenidos mediante el acuerdo. Supongamos por ejemplo que el resultado del regateo es (65,85) para esclavos y amos respectivamente si se parte de una situación inicial que respeta la salvaguarda, y (55,95) si no se respeta. Por otra parte, la situación actual no cooperativa tiene un resultado de (20,60). Para los amos, el acuerdo satisfactorio para los esclavos representa una pérdida de 10 unidades de utilidad con respecto al resultado insatisfactorio. Es de suponer que los amos preferirán el resultado satisfactorio si los costos de mantenimiento del acuerdo insatisfactorio representan una pérdida del al menos 10 unidades de utilidad.

Este ejemplo muestra dos cosas. En *primer* lugar, muestra que, en una situación en la que se violen las condiciones de negociación, los favorecidos por el estado de cosas sólo aceptarán negociar sin tener en cuenta los resultados de esa violación en caso de que el coste de mantenimiento de un acuerdo insatisfactorio sobrepase los beneficios que les reporta ese acuerdo. En *segundo* lugar, muestra que los favorecidos toma-

²³ No se pretende que este supuesto se cumpla necesariamente. Ni siquiera se pretende que sea especialmente realista. Definir un agente racional como aquel que persigue la maximización de la utilidad no significa que sus intereses sean puramente egoístas. Pueden serlo o pueden no serlo. En este momento, eso es indiferente. La cuestión es que si dos agentes o dos grupos de agentes se proponen iniciar una negociación es porque sus intereses ni son idénticos. Y no debe tampoco olvidarse que el juego de regateo es un juego estrictamente competitivo. A pesar esto, puede muy bien suceder que en algún caso alguno o todos los participante en el regateo tengan interés en el bienestar del otro. Pero si esto sucede, este interés no egoísta está incluido en la función de utilidad de cada uno.

rán también en cuenta las probabilidades de que los desfavorecidos decidan romper el acuerdo tan pronto como puedan. Es decir, tomarán en cuenta las probabilidades de que el costo de mantenimiento de los medios coercitivos tenga lugar. Veamos esto en primer lugar.

Para tratar esta cuestión debemos preguntarnos si un agente racional estará dispuesto a buscar las oportunidades de romper un acuerdo insatisfactorio para él. Hay que tener en cuenta que no sólo el mantenimiento de medios coercitivos resulta costoso, sino que también lo es enfrentarse a ellos. Es decir, el acuerdo insatisfactorio resulta costoso para todas las partes. Por ello, es de suponer que, al igual que los beneficiados por el acuerdo insatisfactorio deben plantearse si les compensa afrontar sus gastos de mantenimiento, los perjudicados deben plantearse si les compensa afrontar los costos del enfrentamiento con los medios coercitivos empleados para mantener el acuerdo. Para ello, deberán comparar los beneficios de un acuerdo satisfactorio con tales costos. Sin embargo, esto no es exactamente así. Veamos por qué.

En este sentido es interesante analizar la distinción utilizada por Gauthier entre obediencia plena y restringida. Según Gauthier, "una persona dispuesta a una obediencia plena compara los beneficios que espera de la cooperación en cualesquiera términos con lo que espera de la no cooperación y acepta si los primeros son mayores". En contraste, "una persona dispuesta a una obediencia restringida compara los beneficios que espera de la cooperación con lo que espera de un resultado imparcial y óptimo, y acepta una estrategia conjunta si lo primero se aproxima a lo segundo."²⁴ Gauthier argumenta que un agente racional no estará dispuesto a la obediencia plena sino sólo a una obediencia restringida. La razón es que una persona que mostrara la primera disposición estaría invitando a los demás a aprovecharse de él, puesto que esta dispuesto a aceptar acuerdos tremendamente desventajosos²⁵.

Por tanto, sólo la disposición a la obediencia restringida es un

²⁴ (86, pp. 225-61).

²⁵ Por ejemplo, supongamos que una situación no cooperativa tiene un resultado de (5,6) para los agentes 1 y 2. Si cooperan, pueden conseguir un resultado cuyo valor es 15 y tal que puede repartirse de cualquier modo. Si 1 está dispuesto a una obediencia plena, aceptará cooperar para conseguir un resultado de (6,9) y, supuesto que 2 es un agente racional, no conseguirá más. Esta disposición por parte de 1 mostraría que no es un agente racional.

disposición racional. Esta disposición implica que sólo se aceptará cooperar si el resultado de la cooperación ofrece lo máximo que se puede conseguir teniendo en cuenta una aspiración similar por parte de los demás. Es decir, sólo se aceptará cooperar si el resultado ha sido decidido por un procedimiento imparcial como los discutidos en el capítulo anterior.

Sin embargo, para Gauthier esto no es suficiente. Para él, la disposición racional hacia la obediencia restringida significa algo más. Significa que sólo se aceptarán los acuerdos logrados a través de un proceso de negociación cuyo punto de partida sea satisfactorio.

De nuevo, el motivo es que estar dispuesto a aceptar acuerdos insatisfactorios invita a los demás a negociar con ventaja. Si los esclavos estuvieran dispuestos a respetar acuerdos insatisfactorios, los amos preferirían estos acuerdos. Por ello, supuesto que los esclavos son agentes racionales, no estarán dispuestos a cumplir los acuerdos insatisfactorios. De este modo, los amos se verán forzados a aceptar las condiciones de negociación y los acuerdos basados en ellas, o tendrán que afrontar la creación de medios coercitivos. Dicho de otro modo, el único medio del que disponen los esclavos para forzar a los amos a aceptar un acuerdo satisfactorio es hacerles ver los costos que podrían imponerles en caso contrario. Igualmente, el único medio de los amos para hacer cumplir a los esclavos un acuerdo menos que satisfactorio es disponer de unos medios coercitivos que hagan ver a los esclavos los costos que tendrían que afrontar de no cumplir el acuerdo. Por consiguiente, en caso de darse un acuerdo insatisfactorio, los amos saben que tendrán que emplear tales métodos con seguridad puestos que los esclavos estarán siempre dispuestos a incumplir el acuerdo a la menor oportunidad. La única alternativa de los amos esta en no darles esa oportunidad.

La cuestión de si los favorecidos por el acuerdo insatisfactorio estarán dispuestos a renunciar a estos privilegios con el fin de conseguir un acuerdo satisfactorio depende de si los beneficios suplementarios que les reporta el acuerdo insatisfactorio son suficientes para compensar el costo que supone el empleo de medios coercitivos.

Supongamos que los participante en la negociación pueden calcular estos costos. Por ejemplo, un acuerdo insatisfactorio entre amos y esclavos tendría como resultado (55.95), mientras que el resultado de un a-

acuerdo satisfactorio sería (65,85). Todos saben que de los beneficios que los años extraen del cumplimiento de este acuerdo hay que descontar una cantidad x en concepto del costo de mantenimiento del acuerdo. Supongamos que x vale 11, con lo cual el resultado será (55,84). Ahora bien, esta situación sería subóptima puesto que hay otra situación posible en la que el resultado es mejor para todos. Si cooperan de un modo satisfactorio todos salen ganando.

Sin embargo, hay otras situaciones en las que la justificación de las condiciones de negociación no resulta tan clara:

1- Supongamos ahora que x vale 9. Esto hace que el resultado de la situación sea (55,86). Al contrario de lo que ocurría en el caso anterior, la situación es óptima. Cualquier otra situación, incluida la resultante de un acuerdo satisfactorio, sólo aumentan la utilidad de una de las partes a costa de disminuir la de otra.

Este tipo de situaciones plantea un problema que ya conocemos. En efecto, sabemos que la cooperación sólo puede surgir cuando la situación no cooperativa es subóptima, o dicho de otro modo, cuando la situación no cooperativa tiene la estructura de un juego de intereses mixtos. Del mismo modo, puede decirse que unas condiciones de negociación se aceptarán sólo en el caso de que la cooperación surgida de un acuerdo insatisfactorio sea subóptima. Pero hay casos en los que esto no sucede. Y en estos casos, no parece fácil descubrir por qué un agente racional debe aceptar las condiciones de negociación. Naturalmente, hay que tener en cuenta que un acuerdo insatisfactorio supone unos costos adicionales a los beneficiados por el acuerdo. Pero, aun así, pueden darse casos en los que compense afrontar esos costos. En estos casos, es de esperar que no se respeten las condiciones de negociación²⁴.

²⁴ Puede pensarse que una situación de cooperación insatisfactoria se parece enormemente a una situación no cooperativa. En nuestro ejemplo, si la cooperación es insatisfactoria los años tendrán que emplear métodos coercitivos para obligar a los esclavos a cumplir el acuerdo. Pero, al fin y al cabo, esto es lo que sucedía antes de la cooperación. Los años tenían que emplear métodos coercitivos para mantener la situación. Entonces, ¿qué es lo que se sale ganando con el acuerdo? Sin embargo, ambas situaciones son muy distintas. La situación no cooperativa es subóptima, lo cual significa que todos están interesados en cooperar. Debido a esto, todas las partes pueden ejercer una presión determinada para conseguir un acuerdo que les resulte tan beneficioso como sea posible. Esto es precisamente lo que no sucede en los casos que ahora nos ocupan. En ellos aparece una situación óptima. Esto significa que el grupo o el individuo beneficiado por la situación no tiene ningún motivo para desear cambiarla. Lo único que puede hacer el agente años favorecido es intentar que su oponente encuentre rentable un cambio de

2- Gauthier mismo plantea otro tipo de casos²⁷, en los que el problema es que el acuerdo satisfactorio tiene un resultado aún peor que la no cooperación para una de las partes. Naturalmente, la parte perjudicada por el acuerdo satisfactorio nunca accedería a negociar en esos términos. Por consiguiente, parece que si la otra parte quiere llegar a un acuerdo tendrá que renunciar a exigir que las condiciones de negociación sean respetadas, ya que en este caso son ellas las que imposibilitan la cooperación.

En estos casos Gauthier, siguiendo a Buchanan, admite que ambas partes negocien en base a la situación real no cooperativa. Para hablar de estos casos, Gauthier emplea un término especial, "aquiescencia", con objeto de distinguir la disposición de un agente racional hacia este tipo de situaciones tanto de la obediencia plena como de la restringida. Un agente racional aceptará un acuerdo menos que satisfactorio y respetará el acuerdo en estos casos, aunque siempre estará dispuesto a pedir un replanteamiento de la situación tan pronto como sea posible lograr un acuerdo satisfactorio aceptable por ambas partes.

Estos dos tipos de casos tienen cosas en común. En ambos se plantean problemas para conseguir acuerdos satisfactorios. Y en ambos casos la razón parece ser la misma, a saber, que para una de las partes no resulta racional aceptar las condiciones de negociación y esa parte es precisamente la que puede ejercer una presión mayor para decantar el acuerdo a su favor, e.d., la parte que ocupa una mejor posición en la situación no cooperativa. Por ello, puede parecer sorprendente que Gauthier ignore por completo el primero de estos casos. Esto, sin embargo, no es casual. Obedece a que, para Gauthier, sólo en el segundo caso sería justificable que un agente racional violara las condiciones de negociación. Para Gauthier "la salvaguarda constriñe la posición inicial de regateo hasta el extremo, y sólo hasta el extremo, en que tal constreñimiento es compatible con el resultado cooperativo que proporciona a cada persona la expectativa de una utilidad mayor que la que le ofrece el resultado no cooperativo" (86,p.230). Según parece desprenderse de estas palabras, sólo

la situación, y el único modo en el que esto puede lograrse es imponiéndole tales costos que la situación pase a ser subóptima. Sólo entonces será posible una nueva negociación.

²⁷ (86,pp.227-9)

en el caso de que el respeto a las condiciones de negociación signifique una disminución de la utilidad *por debajo del nivel conseguido con la no cooperación* tendría motivos un agente racional para violarlas. Pero cuando el acuerdo satisfactorio sólo significa una disminución de la utilidad con relación al acuerdo insatisfactorio un agente racional debe seguir prefiriendo el primero.

Es difícil entender el motivo de esta diferencia. La argumentación a favor de la salvaguarda no parece justificarla. Todo el peso de esta argumentación reposa en la distinción ya mencionada entre obediencia plena y restringida, y en mostrar como sólo la disposición a la segunda es una disposición racional. Es indudable que una disposición a la obediencia plena no es racional, en tanto que invitaría a los otros a aprovecharse de los así dispuestos al sentar las bases de la negociación. Al contrario, la actitud racional es hacer saber a los demás que, si el acuerdo no es satisfactorio, los favorecidos por el acuerdo deberán hacer frente a los costos que supone el mantenimiento de un acuerdo que no se esta dispuesto a seguir voluntariamente. Pero de aquí no parece seguirse que un agente racional deba aceptar siempre las condiciones de negociación con la única excepción de que esto suponga obtener unos beneficios aún menores que los que ofrece la no cooperación.

Sin embargo, hay que tener en cuenta que Gauthier cree haber mostrado con su argumento más que eso. Concretamente, lo que él cree haber mostrado es que la disposición hacia la obediencia restringida, que supone la aceptación de las condiciones de negociación, *es la propia de un agente racional sea cual sea la situación en la que este se encuentre*. Es decir, pretende haber mostrado que no sólo es racional para los perjudicados por un acuerdo insatisfactorio el estar sólo dispuestos a la obediencia restringida, sino que lo mismo sirve para los beneficiados por tales acuerdos.

En este sentido la argumentación consta de dos partes. En *primer* lugar, se afirma la racionalidad de no estar dispuesto a más de una obediencia restringida. Tras mostrar que una disposición a la obediencia plena es irracional, se pasa a afirmar que es irracional para cualquier agente. Este primer paso no presenta especiales dificultades.

La *segunda* parte de la argumentación es más problemática. En ella se afirma que si *todos* fueran menos que obedientes en sentido restringido

la cooperación sería imposible. Esto sucede porque entonces nadie estaría dispuesto a aceptar un acuerdo a menos que su resultado le favoreciera de modo parcial o bien mediante una concesión superior a la racional por parte de los otros o bien mediante una violación de las condiciones de negociación que les favoreciera. Ahora bien, ¿qué pasaría si algunas personas fueran menos que obedientes en sentido restringido? Parece que en este caso la cooperación sólo sería posible si otros fueran más que obedientes en sentido restringido. Y esta posibilidad está vetada por el supuesto de igual racionalidad, ya que igual racionalidad implica igual disposición a la obediencia. Por consiguiente, ningún agente racional será menos que restringidamente obediente, lo cual significa que todos aceptarán las condiciones de negociación que hacen que los acuerdos sean satisfactorios.

Gauthier afirma que si alguien es menos que restringidamente obediente la cooperación sólo es posible si alguien lo es más. Sin embargo, tal afirmación es inexacta. Lo que sucede es que si alguien es menos que restringidamente obediente entonces una cooperación estable, basada en un acuerdo satisfactorio, es imposible²⁰. De modo que si se quiere una cooperación basada en un acuerdo satisfactorio, entonces no se debe ser menos que restringidamente obediente. Pero, como ya hemos visto, no siempre es racional respetar las condiciones de negociación. No hay por tanto ningún motivo por el que un agente racional deba querer en todos los casos un acuerdo satisfactorio. Es perfectamente posible la cooperación sin necesidad de violar el supuesto de igual racionalidad.

Si esto es así, entonces es interesante averiguar por qué Gauthier insiste en que es siempre racional aceptar la salvaguarda. La explicación más plausible es que no haya tenido suficientemente en cuenta la distinción entre dos preguntas independientes aunque relacionadas, a saber, 1) en qué circunstancias es racional realizar acuerdos cooperativos y 2) en qué circunstancias es racional mantener los acuerdos. Cuando no se tiene en cuenta esta distinción y, especialmente, cuando se toma la

²⁰ No es necesario ni preciso añadir "a menos que alguien sea más que restringidamente obediente". Un acuerdo insatisfactorio no se caracteriza porque las partes no estén de hecho dispuestas a mantenerlo, sino por ser tal que ningún agente racional estaría dispuesto a mantenerlo. Y puesto que la disposición a más de una obediencia restringida es una disposición irracional, un acuerdo en el que alguna de las partes muestre una obediencia más fuerte que esta es un acuerdo insatisfactorio.

respuesta a la segunda pregunta como respuesta también a la primera, es fácil afirmar que no es racional llegar a acuerdos insatisfactorios.

No obstante, Gauthier reconoce explícitamente esta distinción. Por ejemplo, utilizando como ilustración el caso de los esclavos, reconoce que cuando el poder de los amos es suficiente para forzar a una negociación insatisfactoria y el coste de la resistencia por parte de los esclavos es superior al beneficio que tal resistencia puede suponer, la racionalidad exige que se acepte cooperar en esos términos insatisfactorios (86,p.195). De hecho, en este texto Gauthier emplea el término "aquiescencia" para designar la actitud de los esclavos ante estos acuerdos.

Sin embargo, más adelante parece querer limitar el alcance de este término a los casos en los que los perjudicados por el acuerdo insatisfactorio deben racionalmente llegar a un acuerdo debido a que los beneficiados nunca aceptarían un acuerdo satisfactorio porque esto les supondría un descenso del nivel de utilidad por debajo del alcanzado en la situación no cooperativa, e.d., a los casos pertenecientes al segundo grupo (p.116). En este caso sería racional entrar en la negociación porque de otro modo la cooperación sería imposible. La cooperación sólo puede darse cuando todos se benefician con ella. Por tanto no deben exigirse unas condiciones de negociación cuyo efecto sea que una de las partes se ve perjudicada por la cooperación. Esto es sin duda cierto. Pero es igualmente cierto que, en los casos en los que un acuerdo insatisfactorio resulta rentable, no sólo los favorecidos por el acuerdo no tienen ningún motivo racional para aceptar la salvaguarda, sino que los perjudicados actuarán de un modo racional al entrar en la negociación. Si no lo hacen, no habrá cooperación. Y en este caso, naturalmente, todos salen perdiendo, pero ellos pierden mucho más. La capacidad de resistencia de los peor situados en la situación no cooperativa es mucho menor que la de los mejor situados. Y puesto que saben que los favorecidos no tienen ningún motivo racional para aceptar otras condiciones de negociación, obrarán racionalmente aceptando la cooperación en los términos que estos puedan imponerles. Pueden pensarse que los menos favorecidos pueden forzar a los más favorecidos a una negociación satisfactoria si pueden hacerles creer que su determinación a no aceptar unas condiciones insatisfactorias es superior a la suya de no aceptar un acuerdo satisfactorio. Pero ¿cómo podrían hacerlo? Si ninguno da su brazo a torcer y la cooperación no se



logra, ellos pierden mucho más. Por consiguiente, si son racionales su determinación es menor. En este sentido, el por qué los peor situados tienen que aceptar una negociación basada en un punto de partida insatisfactorio es el mismo que explica por qué los peor situados en la posición inicial de regateo tienen que aceptar un acuerdo cuyo resultado es peor para ellos que para los mejor situados.

En resumen, no es racional mantener acuerdos insatisfactorios. Por ello, en muchas ocasiones un agente racional deseará lograr acuerdos satisfactorios y, para ello, deberá basar sus negociaciones en una posición inicial de regateo que respete la salvaguarda. Sin embargo, hay ocasiones en las que un agente racional encuentra que un acuerdo insatisfactorio le resulta más beneficioso que un acuerdo satisfactorio a pesar del coste que supone su mantenimiento. En tales ocasiones, un agente racional no tendrá ningún motivo para basar sus acuerdos en el respeto a la salvaguarda. Y los perjudicados por el acuerdo insatisfactorio, a pesar de que no resulte racional mantener el acuerdo, encontrarán racional entrar en la negociación y llegar a un acuerdo cooperativo.

Antes de pasar a otra cuestión, queda por contestar una objeción que puede hacerse a este planteamiento. La objeción es la siguiente. Hemos dicho que no es racional mostrarse dispuesto a cumplir un acuerdo insatisfactorio. Más bien al contrario, un agente racional estará dispuesto a romper el acuerdo tan pronto como le sea posible. Por tanto, la parte favorecida por el acuerdo insatisfactorio tendrá que emplear una serie de métodos coercitivos para garantizar el cumplimiento del acuerdo. Pero entonces ¿para qué nos sirve un acuerdo de este tipo? ¿no podrían los que ejercen el poder disponer de esos mismos medios para obligar a los otros a actuar del modo deseado sin necesidad de que medie un acuerdo de ningún tipo?

Esto puede ilustrarse acudiendo de nuevo a el caso de los esclavos. La situación no cooperativa resulta subóptima debido al costo que supone mantener la situación, y también a que los esclavos no trabajan a gusto y, por ello, no rinden todo lo que podrían. Un acuerdo insatisfactorio no supondría una mejora sobre esta situación. Los esclavos (ex-esclavos si se prefiere) servirán a los amos "libremente" sólo si los amos disponen de medios para hacerles servir. Pero esto es precisamente lo que ocurría antes de la cooperación. Tampoco servirán gustosamente. Lo fingi-

rán si se les obliga a ello, pero esto también pasaba antes. Ahora hacen lo que hacen obligados, pero para obligar a alguien a hacer algo no hace falta llegar a un acuerdo con él. Parece que, a pesar de todo, un agente racional, si desea un acuerdo de algún tipo, debe desear un acuerdo satisfactorio, pues un acuerdo de otro tipo es igual que una situación no cooperativa. Esta objeción puede ser entendida de dos modos. En *primer* lugar, puede querer decir que un acuerdo insatisfactorio no es rentable o, al menos, que su resultado, al igual que el resultado de la situación no cooperativa, es subóptimo. Entendida de este modo, la objeción ya ha sido contestada más arriba (ver p.115).

En *segundo* lugar, puede querer decir que una situación en la cual el cumplimiento de un acuerdo ha de ser garantizado por métodos coercitivos no merece ser llamada cooperativa. Sin embargo, al definir una situación cooperativa no se hace mención alguna del método empleado para hacer efectivo el cumplimiento de los acuerdos. Lo único que se requiere es que los jugadores tengan libertad para acordar estrategias conjuntas constringentes destinadas a promover sus intereses comunes²⁷.

Esto nos lleva a la siguiente cuestión que es preciso tratar. Esta cuestión no es otra que aquella con la que se encabezaba este capítulo y para la cual lo dicho aquí ha sido una preparación. Esta cuestión es cuando y por qué es racional mantener los acuerdos y a ella estará dedicado el siguiente capítulo.

²⁷ Aunque esta es la definición habitual de un juego cooperativo (por ejemplo, Luce y Raiffa 57 p.89), algunos autores han propuesto una definición aun más amplia. Por ejemplo, Harsanyi (77c,p.110) define un juego cooperativo como aquel en el que los acuerdos son siempre vinculantes y un juego no cooperativo como aquel en que no lo son. La justificación que se da de esta modificación es que la definición amplia permite analizar los efectos de la variación independiente de las dos variables implicadas en la definición habitual, a saber, la imposibilidad de los compromisos por un lado y la libertad de comunicación entre los jugadores por otro.

CAPITULO VII

LA ESTABILIDAD DE LOS ACUERDOS SATISFACTORIOS

En los capítulos anteriores hemos analizado el proceso de regateo mediante el cual unos agentes racionales llegarán a un acuerdo sobre la realización de uno de los puntos óptimos posibles mediante la utilización de la estrategia conjunta que tiene a ese punto como resultado.

Sin embargo, la existencia de un acuerdo de este tipo no es suficiente. Los acuerdos sólo tienen sentido si son estables, e.d., si los participantes pueden suponer razonablemente que los acuerdos, una vez alcanzados, se mantendrán. Por lo tanto, tenemos que saber que condiciones hacen que los acuerdos sean estables. Como se verá con facilidad, esta pregunta esta estrechamente relacionada con la cuestión acerca de cuando es racional mantener un acuerdo. Esta relación obedece a que, por lo general, un agente racional sólo mantendrá un acuerdo si tiene motivos para esperar que los demás también lo harán.

Como señala Harsanyi (77c,p.110), parece que sólo hay dos formas en las que un acuerdo puede ser estable, a saber, bien porque el acuerdo sea auto-imponible o bien porque sea imponible. Un acuerdo es auto-imponible cuando el resultado que para cada jugador se sigue del acuerdo ofrece un incentivo suficiente para garantizar su cumplimiento. Esto significa que un acuerdo sólo puede ser auto-imponible si corresponde a un punto de equilibrio fuerte, mientras que un punto de equilibrio débil sólo será estable bajo ciertas condiciones especiales¹. Por otro lado, un acuerdo es imponible si las reglas del juego exigen que los jugadores mantengan el acuerdo incluso si les resulta beneficioso romperlo².

¹ El motivo por el que sólo los puntos de equilibrio fuerte garantiza la estabilidad del acuerdo es que, en el caso de un punto de equilibrio débil, un jugador puede romper el acuerdo actuando según una estrategia distinta de la acordada sin que esto signifique para él una pérdida de utilidad (aunque, por otro lado, tampoco le suponga un beneficio romper el acuerdo). Para una discusión más extensa de este punto, ver Harsanyi (77c) pp.124-7.

² Un ejemplo claro de acuerdo auto-imponible es el que presentan los juegos de intereses idénticos. Podemos ver esto utilizando una ligera modificación del ejemplo 1 del capítulo IV.

En el capítulo anterior analizamos la distinción entre acuerdos satisfactorios y acuerdos insatisfactorios. Un acuerdo insatisfactorio se caracteriza por ser tal que ningún jugador racional estaría dispuesto a mantenerlo en ausencia de métodos coercitivos dispuesto a este fin. Por este motivo Gauthier afirma que este tipo de acuerdos no son estables. Sin embargo, si utilizamos el concepto de estabilidad tal y como lo hemos definido en la página anterior, un acuerdo insatisfactorio puede ser estable.

Por supuesto, un acuerdo insatisfactorio no es auto-imponible. Pero puede ser imponible y, por ello, estable. Probablemente, la razón por la que Gauthier piensa que un acuerdo insatisfactorio no puede ser estable es porque para él la estabilidad "no natural", e.d., la que no surge de un punto de equilibrio, sólo puede lograrse si el acuerdo es capaz de ganar la aceptación "de todo corazón" por parte de la totalidad de los participantes" (86,p.230). Este modo de conseguir la estabilidad, e.d., mediante un acuerdo que se "gane el corazón" de los participantes, es sin duda distinto a los dos mencionados anteriormente. Es distinto a lo que Gauthier llama "estabilidad natural", puesto que esta sólo es posible cuando el acuerdo es un punto de equilibrio fuerte, cosa que no sucede en los casos de intereses mixtos. Y, desde luego, es distinta de la estabilidad conseguida cuando los acuerdos son imponibles. Es, por tanto, un tercer tipo de método por el cual un acuerdo puede conseguir

	B_1	B_2
A_1	(15,15)	(5,5)
A_2	(0,0)	(10,10)

Ejemplo 1'

Como dijimos, si ambos jugadores son racionales utilizarán el par de estrategias A_1 y B_1 . Naturalmente, en este caso no media ningún proceso de regateo para acordar las estrategias que cada jugador utilizará. Pero lo importante es observar que el jugador 1 sólo utilizará la estrategia A_1 si espera que el jugador 2 utilice a su vez B_1 , ya que en caso contrario utilizará A_2 . Lo mismo puede decirse respecto al jugador 2. Ahora bien, cada uno tiene buenos motivos para suponer que el otro utilizará esa estrategia, puesto que obrar de otra manera supondría un descenso de la utilidad. Técnicamente hablando, lo que sucede es que en los juegos de intereses idénticos unos jugadores racionales siempre actuarán según la estrategia que corresponde a un punto de equilibrio fuerte. Por ello, cada jugador esperará que el otro utilice la estrategia señalada en tanto que supone que el otro es un jugador racional.

una estabilidad que Gauthier llama "artificial". Quizá el mejor medio de averiguar en que consiste este método sea fijarnos en los acuerdos que, según Gauthier, gozan de una estabilidad así conseguida, e.d., los acuerdos satisfactorios.

Volvamos para ello al caso familiar de los amos y los esclavos. Supongamos la matriz de pagos siguiente

	B_1	B_2
A_1	(65,85)	(15,110)
A_2	(90,40)	(20,60)

Ejemplo 7

Esclavos y amos (jugadores 1 y 2 respectivamente) han realizado un proceso de regateo, mediante el cual han acordado el punto (65,85). En este momento, tal y como refleja la matriz, la situación está del modo siguiente. Si ambos mantienen el acuerdo y actúan siguiendo las estrategias cooperativas (A_1, B_1), el resultado será en efecto (65,85). Si ninguno de los dos cumple el acuerdo y siguen las estrategias no cooperativas (A_2, B_2), entonces se seguirá el resultado no cooperativo (20,60) huyendo del cual nuestros jugadores se han embarcado en el proceso de negociación. Ahora bien, ¿qué pasaría si el acuerdo fuera respetado por una de las partes y no por la otra? Supongamos para simplificar que sólo hay un modo de incumplir el trato, a saber, utilizando la estrategia no cooperativa en vez de la cooperativa, e.d., que no se puede "cooperar hasta cierto punto", o, si se prefiere, que todo incumplimiento del acuerdo por una de las partes tendría el mismo resultado. Si los esclavos cumplen su parte del trato mientras que los amos rompen el acuerdo, el resultado será (15,110). Intuitivamente, podemos interpretar este resultado como indicando que los esclavos pierden 5 unidades de utilidad por debajo del resultado no cooperativo debido al trabajo adicional que tienen que realizar siguiendo los términos del acuerdo. Por su parte, la ganancia de los amos sobre el resultado cooperativo se debe a que no dan a los esclavos lo acordado. Paralelamente, si son los esclavos los que incumplen el trato, el resultado será (90,40). Por último, y dado que queremos saber que tipo de estabilidad tiene un acuerdo satisfactorio al margen de la

concedida por el posible carácter imponible del acuerdo, supondremos que el acuerdo no tiene este carácter. En esta situación, no hay nada que obligue a los jugadores a respetar el acuerdo. Por tanto, ninguno de los jugadores tiene ninguna garantía "externa" de que el otro vaya a cumplir su parte. ¿Cuál sería la conducta de un jugador racional?. Sigamos el razonamiento del jugador 1. Si el jugador 2, utiliza la estrategia cooperativa B_1 , entonces lo mejor que puede hacer es no cooperar. Si el jugador 2 utiliza la estrategia no cooperativa B_2 , de nuevo lo mejor que puede hacer es no cooperar. Y si el jugador 2 utiliza una estrategia mixta cualquiera, otra vez hará mejor 1 no cooperando. El razonamiento de 2 es estrictamente paralelo.

Es decir, la estrategia no cooperativa A_2 es dominante en sentido fuerte tanto con respecto a A_1 como con respecto a cualquier estrategia mixta posible, pues es la única mejor respuesta de 1 con respecto a cualquier posible estrategia de 2, y lo mismo puede decirse de B_2 con respecto a B_1 . Por ello el par de estrategias no cooperativas (A_2, B_2) corresponde al único punto de equilibrio fuerte. Además, en este caso este es el único punto de equilibrio. Por tanto, si el acuerdo no es imponible y ambos jugadores son racionales, ninguno cooperará.

Supongamos que la situación es ligeramente distinta. Ahora, si una de las partes rompe el acuerdo, la parte que lo cumple obtiene el mismo resultado que tendría para el la situación en la que ninguno cooperara. Es decir

	B_1	B_2
A_1	(65, 85)	(20, 110)
A_2	(70, 60)	(20, 60)

Ejemplo 7'

En este caso, la estrategia A_2 también es el elemento maximal del conjunto de estrategias posibles de 1 bajo la relación de dominación fuerte, así como B_2 lo es en el caso del jugador 2. Por tanto también en este caso el par de estrategias no cooperativas (A_2, B_2) corresponden al único punto de equilibrio fuerte del juego. La novedad que presenta este caso con respecto al anterior es que en esta ocasión existe otros dos puntos de equilibrio, si bien de equilibrio débil, a saber los correspondientes

a las combinaciones de estrategias (A_1, B_2) y (A_2, B_1) . Esto sucede porque cada uno de los jugadores, en el supuesto de que el otro utilizará la estrategia no cooperativa, sería indiferente entre el uso de cualquiera de las dos estrategias puras disponibles, ya que en ambos casos conseguiría el mismo resultado³.

Al igual que el caso anterior, el acuerdo logrado en esta situación tampoco resulta estable, ya que para cada jugador la estrategia no cooperativa es la única mejor respuesta para cualquier posible estrategia del otro, exceptuando que el otro utilice la estrategia no cooperativa, en cuyo caso la no cooperación es la mejor respuesta. Pero puesto que cada jugador sabe que para el otro la no cooperación es la única estrategia dominante en sentido fuerte, sabe que será esta la utilizada. El resultado es que ninguno de los dos mantendrá el acuerdo.

Podemos modificar nuestro ejemplo de muchas maneras y el resultado siempre será el mismo, a saber, que el par de estrategias cooperativas nunca corresponden a un punto de equilibrio fuerte. La razón es que en todos los casos ocurre a) que si el otro no mantiene el acuerdo, entonces o bien se gana más no cooperando tampoco o, al menos, se gana lo mismo de un modo que de otro y b) que si el otro mantiene el acuerdo, entonces se gana también más no cooperando. Ahora bien, en ausencia de medios que hagan imponible un acuerdo, sólo los puntos de equilibrio fuerte son estables. Puede darse el caso de que el par de estrategias cooperativas constituyan un punto de equilibrio débil. Pero esto no proporciona estabilidad al acuerdo. Si en el juego existe además un punto de equilibrio

³ Puede pensarse que es posible, e incluso probable, que un jugador no sea indiferente entre estos dos casos, pues si el otro no colabora, él preferirá penalizarlo no colaborando tampoco disminuyendo así la utilidad del contrario. Incluso puede pensarse que todo jugador racional obraría de este modo y se lo haría saber al contrario, pues así disminuiría el incentivo que supone el otro la defección. Sin duda esto es cierto, pero irrelevante en el caso presente, ya que estamos presentando el juego como si los jugadores tuvieran que decidir sus estrategias independiente y simultáneamente (de adelante discutiremos el problema planteando el juego mediante movimientos sucesivos).

También puede pensarse que es probable que un jugador no sea indiferente por otro motivo. Es probable que un jugador razonara del siguiente modo: "si el otro no coopera y yo sí, a pesar de que yo reciba lo mismo tanto si coopero como si no, la utilidad no será la misma en ambos casos, pues en el primero el conseguiré una ganancia enorme a mi costa, la cual me pondrá tan furioso y me llenará de tal modo de justa indignación que la utilidad se vera considerablemente reducida". También esto es probablemente cierto, al menos en muchos casos. Sin embargo, en este caso (y por lo general, a menos que se indique explícitamente lo contrario) expresamos los distintos pagos de los jugadores en términos de utilidad, por lo cual esa consideración y cualquier otra de ese tipo ya está incluida en la función de utilidad de cada jugador.

fuerte, se elegirán las estrategias que correspondan a este último. Si todas las posibles combinaciones de estrategias son puntos de equilibrio débil, entonces el único punto de equilibrio estable será el punto $(1/2A_1 + 1/2A_2, 1/2B_1 + 1/2B_2)^*$.

Lo que esto demuestra es que el resultado de un acuerdo cooperativo no es auto-imponible, por lo cual carece de la estabilidad que Gauthier llama "natural". Esto ocurre tanto si el acuerdo es satisfactorio como si es insatisfactorio. Si suponemos que no existen reglas de juego que hagan el acuerdo imponible, es difícil ver porque el acuerdo puede ser estable. La característica de un acuerdo que, según Gauthier, le hace estable a pesar de no ser imponible ni auto-imponible, como ya dijimos, ha de residir en aquello que distingue un acuerdo satisfactorio de uno insatisfactorio, pues para Gauthier sólo los primeros son estables en este sentido. Pero, ¿qué es lo que puede hacer que un agente racional este dispuesto a cumplir con un acuerdo satisfactorio? Desde luego, un agente racional preferirá que todos cumplan el acuerdo a que ninguno lo haga. Pero esto sucede tanto con los acuerdos satisfactorios como con los que no lo son. Pero esto no significa que este dispuesto a cumplir el acuerdo supuesto que los demás lo cumplan. Significa que esta dispuesto a cumplir su parte *si esto es necesario para asegurar que los demás cumplan la suya*. Esa es la única diferencia entra la actitud de un agente racional hacia los acuerdos satisfactorios y hacia los que no lo son. Pues en el caso de esos últimos, y por los motivos expuestos en el capítulo anterior, un agente racional estará dispuesto a cumplir un acuerdo en tanto que eso es necesario para asegurar que los demás cumplan la suya y además en tanto que los otros tengan el poder suficiente como para obligarles a ello.

Puede pensarse que el planteamiento que hemos hecho respecto a la no estabilidad de los acuerdos satisfactorios encierra una trampa, en tanto que, al presentar la situación de juego en forma normal, lo cual supone que cada agente tiene que decidir la estrategia a utilizar de forma independiente (e.d., simultáneamente a la elección de los demás y sin

* En este caso, cada jugador sería indiferente entre cada una de sus estrategias. Por consiguiente, cada jugador se comportaría como si estuviera utilizando su estrategia centroide de equilibrio (*centroid equilibrium strategy*), e.d., la estrategia mixta que asigna igual probabilidad a todas las estrategias puras. (Ver Harsanyi, 77c, pp.109, 125-6 y 279).



conocer por tanto los movimientos del contrario), por adelantado y de una forma definitiva antes de empezar el juego, hemos eliminado la posibilidad de que el acuerdo sea estable por el motivo explicado de que cada uno cumplirá el acuerdo como condición necesaria para garantizar que los demás también lo cumplen.

Por lo general, las cosas no suceden así. Para empezar, uno habitualmente no tiene que decidir de antemano su estrategia de tal forma que no pueda cambiarla. Más bien lo que ocurre es que uno cambia de estrategia a lo largo del desarrollo del juego y, lo que es aun más importante, cambia su propia estrategia según sea la que los demás están utilizando. Por eso, las cosas pueden suceder de tal modo que un acuerdo satisfactorio resulte estable. Si uno está dispuesto a cooperar sólo si los demás cooperan y si su propia cooperación resulta necesaria para mantener la conducta cooperativa de los demás, uno, por ejemplo, puede empezar el juego dando un primer paso cooperativo. Si los demás no cooperan, entonces en la siguiente jugada se puede dejar de cooperar. Si, por el contrario, los demás cooperan, uno vuelve a cooperar, al menos si se quiere que los demás, a su vez vuelvan a cooperar.

Sin desviarnos de la forma normal, esta situación puede representarse del modo siguiente. Supongamos que los jugadores se ven repetidamente ante el dilema de cooperar o no cooperar. Para decidir en cada caso, cuentan con la información de los resultados que se seguirán de una combinación de estrategias cualquiera en una matriz de pagos. Al final de cada jugada, los participantes reciben el pago obtenido. La diferencia fundamental es que ahora el resultado del juego para cada jugador será la suma de lo obtenido en cada jugada singular. Aunque en cada una de estas jugadas supongamos que los jugadores tienen que decidir de forma simultánea e independiente, cada uno sabrá lo que los demás han hecho en la jugada anterior y podrán realizar su elección a la vista de esta información. Supongamos que un jugador, digamos el jugador 1, en un momento dado realiza la elección de la estrategia cooperativa A_1 . El jugador 2 sabe que la jugada le será más provechosa si el no coopera y utiliza B_2 . Pero también sabe que esto, con toda probabilidad, tendrá como resultado que, en la siguiente jugada, 1 utilice la estrategia no cooperativa A_2 , con lo cual el a su vez se verá llevado a usar B_2 , lo cual resulta perjudicial para ambos. Por tanto, es presumible que se alcance de un modo u otro un

patrón de conducta cooperativa beneficioso tal que ninguno quiera arriesgarse a ponerlo en peligro desviándose a la no cooperación, ya que las ventajas de esta a corto plazo no compensan el perjuicio de la no cooperación a largo plazo. Este planteamiento, que parece responder a la realidad en muchos casos, parece que da estabilidad al acuerdo cooperativo.

Sin embargo, esto no es totalmente cierto. Supongamos que todos saben que en diez ocasiones van a encontrarse en la situación de decidir si cooperar o no cooperar. La última de estas jugadas será considerada por los jugadores como si en realidad fuera un juego de un único movimiento. Por lo tanto, como su comportamiento no repercutirá en el otro, porque no habrá una ocasión posterior, en esta última jugada el resultado será el correspondiente al par de estrategias no cooperativas (A_2, B_2). Puesto que todos saben que en cualquier caso este será el resultado de la última jugada, la penúltima será considerada como si fuera en realidad la última, ya que, ocurra lo que ocurra en esta jugada, la jugada siguiente se desarrollará de modo no cooperativo. Este mismo razonamiento puede llevarse hacia atrás hasta llegar a la primera jugada². Por lo tanto, el acuerdo no será estable.

Esto sirve para cualquier caso en el que el juego vaya a ser jugado en un número n finito de pasos y los jugadores conozcan ese número. Podemos suponer que la situación se planteará un número indeterminado de veces y que los jugadores sólo sabrán qué jugada es la última cuando se enfrenten con ella. En este caso el razonamiento anterior no puede tener lugar. Es posible plantear que en este caso el acuerdo cooperativo sería estable. Pero, de nuevo, esto no ocurre necesariamente en todos los casos.

En esas condiciones, el acuerdo sería estable si el infractor puede ser descubierto en cada caso y "penalizado" con la no cooperación de los demás. Podemos suponer que esto sucede en los acuerdos entre dos jugadores. Pero cuando el acuerdo involucra a más personas la situación no está tan clara. En los juegos n -personales pueden darse dos casos. Puede suceder, en primer lugar, que el resultado cooperativo dependa de la participación de todos. Un ejemplo podría ser el siguiente. Cinco personas quieren colaborar para hacer una salsa mayonesa. El acuerdo al que

² Para una demostración formal, ver Luce y Raiffa (37) pp. 99 y ss.

llegan consiste en que 1 pone el aceite, 2 los huevos, 3 la sal, 4 el limón y 5 realiza la mezcla. Si uno de ellos no cumple con su parte del acuerdo, no hay salsa (al menos no hay una salsa que deseen utilizar). Al igual que sucede en los casos anteriores, todos prefieren que se produzca el resultado del acuerdo a que no se produzca. La diferencia está en que antes si uno no colaboraba y los demás si, el que no cumplía el acuerdo salía ganando, lo cual era de por sí un incentivo para no mantener el acuerdo. Esto no ocurre en este caso. Ahora nadie sale beneficiado por la ruptura del acuerdo, ya que si uno no colabora el resultado para todos es el mismo que si nadie lo hiciera. El único caso en el que uno no mantendría el acuerdo es si los demás tampoco lo hacen. Pero, al menos en tanto los demás cumplan su parte, todos tienen un incentivo para cumplir la suya y ninguno para no hacerlo.

Sin embargo, este tipo de situaciones estaban ya fuera de cuestión, pues son un caso de acuerdo auto-imponible. De hecho, son un caso de intereses idénticos.

Puede darse, por otro lado, un segundo caso que se caracteriza por que el resultado del acuerdo se produce aunque no todos cumplan su parte siempre y cuando colabore un número suficiente. Cuando esto sucede, pueden ocurrir dos cosas. Puede que por las características del caso no sea posible, o, al menos, no sea fácil, localizar al infractor. Y puede ocurrir que, a pesar de que pueda ser localizado, no pueda ser penalizado con la no cooperación de los demás. Esto último puede suceder por varios motivos, de los cuales el más importante es que los demás sigan encontrando ventajoso cumplir con su parte y que el infractor no pueda ser excluido de los beneficios de la cooperación. El ejemplo paradigmático de este tipo de casos son los bienes públicos.

Los bienes públicos, en su mayoría, son unos beneficios tales que resulta imposible privar a alguien de su disfrute. Piénsese, por ejemplo, en la defensa nacional, la descontaminación del aire o el alumbrado público. Todos estos bienes son costeados por los ciudadanos. Todos prefieren que todos paguen su parte a que ninguno lo haga, pero también todos prefieren en primer lugar que los demás paguen y no pagar uno mismo. Esto supone un beneficio mayor. Puesto que el resultado sigue produciéndose, uno sigue beneficiándose de lo producido sin tener que cumplir su parte. Cuando sucede, como es habitual, que los implicados son numerosos,

en ausencia de medidas que garanticen el cumplimiento del acuerdo, entre las que se cuentan medidas para localizar a los infractores, es difícil saber quién esta y quién no esta cumpliendo el acuerdo. En comunidades reducidas la posibilidad de conocer este dato siempre es mayor. Pero incluso si damos por garantizado que el que no cumple el acuerdo será descubierto, sigue siendo imposible excluirle del disfrute del bien producido. En consecuencia, en todos estos casos es posible ser un aprovechado y es racional serlo.

Por todo ello, en estos casos el acuerdo sigue siendo inestable. Cada uno de los jugadores razonará así. Si los demás, o un número suficiente de los demás, no cumplen el acuerdo, entonces yo haré mejor no cumpliéndolo tampoco. Y si los demás, o un número suficiente de los demás, lo mantienen, también obtendré más beneficios no cumpliéndolo.

En resumen, dijimos en el capítulo anterior que, por lo general, un agente racional no mantendrá acuerdos insatisfactorios, por lo cual estos resultan inestables. En este capítulo hemos visto como no siempre es racional mantener los acuerdos satisfactorios si estos no son imponibles. Concretamente, no es racional mantenerlos a) cuando no es previsible que vuelva a darse una situación de colaboración entre los mismos agentes, b) cuando, aunque la oportunidad de cooperar va a presentarse repetidas veces entre los mismos jugadores, la repetición se dará un número finito y definido de veces conocido por los agentes, c) cuando es improbable que el infractor sea reconocido y d) cuando, a pesar de que la probabilidad de reconocimiento sea elevada, los demás no pueden penalizar la no colaboración evitando que los infractores disfruten del producto de la colaboración. En todos estos casos el acuerdo será inestable.

Esto es más importante en tanto que estas condiciones se dan en la gran mayoría de los casos. Parece por tanto que es preciso introducir medidas que hagan que los acuerdos se cumplan, e.d., hacer que los acuerdos sean imponibles. Al análisis de estas medidas dedicaremos el siguiente capítulo. Pero antes nos detendremos a considerar porque Gauthier insiste en la estabilidad de los acuerdos satisfactorios.

Gauthier, siguiendo a Hobbes, intenta demostrar la racionalidad de

cumplir con los acuerdos⁴, y por tanto su estabilidad, cambiando la pregunta por la de si es racional tener una disposición hacia el cumplimiento de los acuerdos. El motivo de este cambio es el reconocimiento del propio Gauthier de que no es posible mostrar que el cumplimiento individual de un acuerdo concreto sea siempre racional. En este sentido, puede que sea racional cumplir los acuerdos, a pesar de que eso nos perjudique, si es posible mostrar que la disposición a cumplirlos es beneficiosa.

Gauthier contesta de forma afirmativa a esta nueva pregunta. La razón, en resumen, es simple. A saber, si los demás conocen mi disposición a mantener los acuerdos, estarán a su vez dispuestos a llegar conmigo a tratos beneficiosos para todos. Si, por el contrario, saben que mi disposición es contraria al mantenimiento de los acuerdos, se guardarán mucho de cooperar conmigo, lo cual, dados los indudables beneficios de la cooperación sobre la no cooperación, será a la larga desventajoso para mí. Por ello, un agente racional tendrá una disposición clara a cumplir los acuerdos, incluso cuando no resulte inmediatamente beneficioso hacerlo.

Sin embargo, Gauthier es plenamente consciente de que este argumento sólo es válido si suponemos que somos transparentes, e.d., que no podemos ocultar a la vista de los demás nuestras disposiciones. Esto no es sino una nueva presentación de las condiciones b y c presentadas más arriba. Pero, como ya dijimos, estas condiciones se cumplen casi siempre. No obstante, esto no parece preocupar a Gauthier, quien, al tiempo de reconocer que la transparencia no es habitual, argumenta que tampoco es necesaria. Basta con que seamos translúcidos. Esto quiere decir que, si bien no somos transparentes, tampoco somos totalmente opacos, en cuyo caso sería imposible averiguar la disposición de cada uno de nosotros. Ser translúcido significa que, si bien los demás no pueden conocer con certeza nuestra disposición, pueden de algún modo "intuirla". Desgraciadamente, Gauthier no nos explica como "intuir" las disposiciones de los demás, ni tampoco muestra que en realidad seamos translúcidos y estas intuiciones se den de un modo u otro. Esto último es sin duda difícil de mostrar. Es cierto que a veces alguien nos da mala espina. Pero, en pri-

⁴ En adelante, y a menos que se indique expresamente lo contrario, siempre que se hable de acuerdos deberá entenderse que nos estamos refiriendo a acuerdos satisfactorios.

mer lugar, esto no ocurre siempre y, en segundo lugar, solemos equivocarnos con considerable frecuencia. En cualquier caso, la translucidez sería un modo más, aunque no muy fiable, de anular la condición c. La única diferencia es que, de mostrarse que se posee esta propiedad, los casos en los que la condición c se cumple disminuirían.

Además, aun en el caso de que fuéramos translúcidos, la cuestión seguiría sin resolverse, pues la aparición de la condición c no garantiza por sí sola la racionalidad de mantener los acuerdos. El motivo es que, como se recordará, hay situaciones en las que es racional incumplir los acuerdos aunque se tenga la certeza de que los demás van a cumplir su parte, a saber, cuando se cumplen el resto de las condiciones.

Por todo ello, y a falta de demostraciones posteriores, no podemos decir que el argumento de Gauthier sea muy convincente⁷. Más bien al contrario, podemos afirmar que el intento de Gauthier de mostrar la estabilidad de los acuerdos satisfactorios no resulta satisfactorio. Los acuerdos sólo son estables si son auto-imponibles o si son imponibles. No hay una tercera vía. Y puesto que los casos de intereses mixtos, que son en los que con propiedad puede hablarse de cooperación y no cooperación, no son auto-imponibles, debemos analizar los mecanismos que hacen que un acuerdo sea imponible.

⁷ Al final del libro, concretamente en el último capítulo, Gauthier ofrece otro argumento, basado esta vez en la naturaleza humana. Según Gauthier, el hombre es un ser social por naturaleza. Esto no es nuevo, pero sí lo es más su modo de entender la sociabilidad humana. Esta no consiste sólo en que el hombre necesite de los demás para vivir o, cuando menos, para vivir mejor, cosa que queda reconocida desde el mismo momento en que se habla de las ventajas de la cooperación sobre la no cooperación, sino en que el hombre muestra una predilección en la cooperación y la participación en tareas colectivas en sí mismas. Es decir, a las ventajas de la cooperación habría que añadir una ventaja suplementaria obtenida por el sólo hecho de la cooperación. Sobre la verdad de esta afirmación habría mucho que decir a favor y en contra. Por ejemplo, Hobbes mantiene justo lo contrario. Pero, en cualquier caso, esto es irrelevante por el motivo mencionado ya en otras ocasiones, a saber, que sea cual sea su origen y fundamento de las preferencias de los individuos, estas se encuentran ya incluidas en su función de utilidad y no pueden contarse de nuevo como algo aparte, ya que esto supondría que se contarían dos veces. Además, este asunto quedará mejor tratado cuando analicemos la función de ciertos factores, tales como la simpatía, en el mantenimiento de la cooperación.

PARTE TERCERA

COMO HACER LOS ACUERDOS IMPONIBLES

"Supongamos que cada uno está dispuesto a hacer lo que será mejor para sí mismo, o para su familia, o para aquellos a los que ama. Hay entonces un problema práctico. A menos que algo cambie, el resultado real será peor para todos. Este problema es una de las principales razones por la que necesitamos algo más que *laissez-faire economics*, por la que necesitamos tanto la política como la moralidad"¹.

Este texto de Parfit resume de modo preciso todo lo que hemos dicho en el capítulo VII. En ausencia de mecanismos que hagan imponibles los acuerdos, y supuesta nuestra concepción de un agente racional, según la cual este utilizará siempre la estrategia que maximice su utilidad individual, la situación siempre se resolverá de un modo no cooperativo. Y este resultado es subóptimo. Todos ganaríamos más si se respetara el acuerdo y el juego se desarrollara cooperativamente. El objeto de este capítulo es el análisis de las distintas soluciones que pueden ofrecerse al tipo de situaciones comúnmente conocidas como "dilema del prisionero".

Las posibles soluciones, así como su caracterización general, aparecen en un cuadro elaborado por Parfit² y que reproduciremos a continuación, ya que, debido a su claridad y exhaustividad, será de gran utilidad en toda la discusión siguiente. Conviene por ello tenerlo siempre presente.

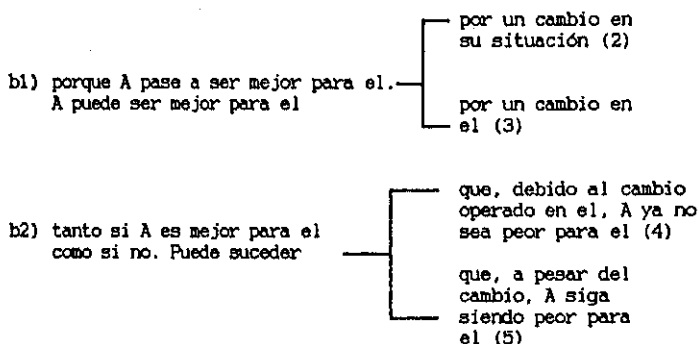
Llamemos "auto-benefica" (B) a la estrategia no cooperativa y altruista (A) a la cooperativa. El motivo de utilizar estos nombres es claro. En un juego en el que los acuerdos no son imponibles, un agente racional que busca maximizar su utilidad empleará la estrategia no cooperativa. Por este motivo nos referiremos a estos juegos como juegos no-cooperativos. Utilizar la estrategia cooperativa en esas condiciones sería tanto como favorecer al contrario a costa de uno mismo, por lo que podríamos calificar de altruista esta elección. Si todos utilizan B, el resultado será subóptimo. Si, por el contrario, todos utilizan A, el problema quedará resuelto. El cuadro que se presenta a continuación refleja los distintos motivos por los que la solución puede ser alcanzada.

¹ Parfit (196), p.62

² Parfit (196), p.63 y (70), p.342

Cada uno puede hacer A

- a) porque B resulte imposible (1)
- b) porque adquiriera una disposición a realizar A. Un agente puede adquirir esta disposición



En este cuadro aparecen numeradas del 1 al 5 los distintos modos en que puede solucionarse el dilema. Dicho de otro modo, estos son los distintos modos que, haciendo que los acuerdos adquirieran estabilidad, transforman un juego no cooperativo en uno cooperativo.

Estas distintas soluciones pueden agruparse en dos categorías. Por un lado están las soluciones políticas, que resuelven el dilema a través de la introducción de un cambio en la situación en la que se encuentra el individuo. A este grupo pertenecen las soluciones 1 y 2. Por otro lado esta el grupo de las soluciones psicológicas, al que pertenecen el resto de las alternativas, y que toman este nombre debido a que dan estabilidad a los acuerdos debido a un cambio producido en el interior del individuo³.

³ Esta división en dos grupos de las razones por las que alguien puede adoptar una estrategia altruista no es nueva. Por el contrario, responde a lo que clásicamente se consideraban sanciones externas e internas. Por ejemplo, Bentham, en un intento de explicar porque alguien puede actuar de un modo que contribuya a la felicidad ajena, ofrece una lista de cinco tipos de sanciones, de entre las cuales las 4 primeras (física, político, popular y religiosa) serían sanciones externas, mientras que la última, la sanción social o simpática, sería externa. De modo similar, Sidgwick habla de sanciones externas e internas para tratar de explicar la coincidencia deber/interés (Sidgwick, 81, p.164). La misma división puede encontrarse en Hill (1974, capítulo III). El esquema clásico presentado por estos autores coincide con el de Parfit no sólo en esta división en dos grupos, sino también en algunas de las alternativas ofrecidas. Otros, como Hobbes, sólo admiten las soluciones políticas.

Las soluciones políticas presentan sobre las psicológicas la ventaja de ser más fáciles de introducir. Pensemos en el caso del dilema n-personal presentado por la contribución al bien público mediante los impuestos. La solución 1 supondría disponer de medios que hicieran imposible la evasión de impuestos, por ejemplo mediante algún tipo de control de ingresos. La solución 2 pasaría por el establecimiento de un sistema de castigos y recompensas que hiciera beneficiosa individualmente la contribución. Estos sistemas son fáciles de introducir en el sentido de que pueden introducirse de una vez por todas mediante la promulgación de una ley. Esto traería en consecuencia la ventaja adicional de que sería fácil que todos conocieran de inmediato la existencia de estas medidas, lo cual haría que cada uno a) conociera que, en la nueva situación, ya no le es posible o no le es beneficiosa la no cooperación y b) supiera que todos los demás están igualmente informados del cambio de la situación y por tanto, también actuarán de forma cooperativa.

Sin embargo, las soluciones políticas cuentan también con no pocos inconvenientes. En el caso de la solución 1, el principal problema es que es, en muchos casos, sencillamente inaplicable. Piénsese por ejemplo en el caso de los soldados presentado por Parfit. Cada uno se enfrenta con la elección de desertar o mantenerse en su puesto. Si los demás desertan, lo mejor que puede hacer es desertar también. Si los demás se mantienen en su puesto, entonces también es mejor desertar. Puede haber distintos modos de convertir esta situación en cooperativa, pero parece que la solución 1 es difícilmente aplicable. Siempre se puede atar a los soldados a sus puestos, o romperles un tobillo. Entonces no podrán huir. Pero tampoco podrán avanzar.

La solución 2 tampoco es aplicable en todos los casos. Pensemos en la práctica de ayudar al prójimo. Si nadie ayuda a los demás, lo mejor será que yo tampoco lo haga y si los demás lo hacen también es mejor para mí si no lo hago. Pero de esta forma, todos salimos perdiendo. A veces es posible utilizar 2 como solución, pero no siempre. Porque para eso debería disponerse de un medio para saber en todos los casos cuando alguien ha ayudado a otro y cuando no. Naturalmente, no bastaría con la palabra de ninguno de los implicados. Y, a menos que dispongamos de una especie de vigilante personal para cada uno, esto parece en muchos casos difícil de averiguar.

La única desventaja de las soluciones políticas no es que no sean universalmente aplicables. Incluso en los casos en los que pueden emplearse (que, dicho sea de paso, son bastantes) no constituyen una buena solución. El motivo es que introducir estas soluciones resulta tremendamente costoso. Si se emplearan en todos los casos, el costo de mantenimiento del sistema coercitivo y del sistema de recompensas sería tan grande que probablemente saldríamos perdiendo aun más que con la no cooperación. En el mejor de los casos, la situación alcanzada mediante estos mecanismos no será óptima. Si podemos conseguir algún otro medio que nos ahorre el costo supuesto por estas soluciones, todos saldremos ganando.

Además, estas soluciones tienen desventajas notables. Por ejemplo, tal y como apunta Rawls⁴, la creación de este sistema cuenta con el inconveniente de ser un peligro para la libertad individual, existiendo siempre el riesgo de una interferencia indebida en los asuntos privados. Esto sin contar con que la aplicación universal de estas medidas sin duda sería contemplada como un "no dejar vivir" por parte del estado, y crearía una sensación tan angustiosa de falta de libertad que sus posibles beneficios se verían considerablemente mermados.

Por todo ello, parece que estas soluciones deben ser consideradas como un último recurso que debe aplicarse en ausencia de un método mejor. De todas formas, y en tanto que su existencia da estabilidad a la cooperación, en ausencia de otra solución, será racional adoptarlas siempre y cuando sus desventajas no hagan que la situación resultante sea aun peor que la no cooperación.

Las alternativas numeradas del 3 al 5 son soluciones psicológicas. Todas ellas solucionan el dilema, e.d., hacen que cada agente seleccione la estrategia A, debido a un cambio producido en el propio agente. Estas soluciones son soluciones morales cuando el cambio operado en el agente no es un cambio de carácter específico sino de tipo general. Al hablar de "cambio de carácter específico" nos referimos a cambios que sólo representa una solución para un dilema concreto. Un caso de cambio específico sería el que se produciría en el ejemplo de los soldados (p.139) si estos desarrollaran una tendencia compulsiva a obedecer las órdenes de los generales, o si adquirieran un odio hacia el enemigo que les hiciera desear

⁴ Rawls (77) p.240

su muerte aun a costa de arriesgar seriamente su propia vida. Estos cambios producidos en los soldados harían que en el caso concreto de la elección entre desertar y permanecer en sus puestos todos eligieran esto último. Pero si estos mismos soldados se encontrarán involucrados en una situación distinta, por ejemplo si se encontrarán en el dilema de pagar o no impuestos, la disposición adquirida no serviría de nada y ellos volverían a elegir la no cooperación.

Las soluciones morales son siempre de carácter general, y consisten en unos cambios en el agente que son operativos en una amplia gama de situaciones⁴. Estos cambios de carácter moral producidos en el agente pueden dar lugar a las soluciones numeradas del 3 al 5. Si consideramos estas soluciones veremos que se dividen en dos grupos. Por un lado están las soluciones 3 y 4. En ellas, la estrategia A deja de ser peor para el agente. La diferencia entre ellas esta en que en la solución 3 el motivo por el que el agente esta dispuesto a seleccionar la estrategia A es por que A es ahora mejor para él. Sin embargo, en la solución 4, el agente actúa motivado por el cambio moral operado en él y el hecho de que A, debido a este cambio, ya no sea peor para él, es un simple efecto colateral. La solución 5, por otro lado, tiene lugar cuando el agente elige A a pesar de que esta elección sigue siendo peor para él.

Las soluciones al dilema se dividen pues en dos grupos. Por un lado aparecen las situaciones en las que la elección de la estrategia A, por diversos motivos, ya no es peor para el agente. A este grupo pertenecen las soluciones numerados de 1 a 4. Por otro lado está una situación en la que la elección de A sigue siendo peor para el agente. Esta es la solución 5.

Lo que tienen en común las soluciones pertenecientes al primer grupo es que representan situaciones en las cuales la función de utilidad de los agentes se transforma de tal modo que el dilema deja de ser tal. Si representamos estas situaciones de juego en su forma normal veremos que la matriz de pagos se altera haciendo que la elección que resulta mejor para el agente sea al mismo tiempo la que resulta más beneficiosa

⁴ Estos cambios, y por consiguiente, las soluciones morales, son diversos. A pesar del interés que el estudio de estos cambios y de sus ventajas relativas tiene por sí mismo, aquí es suficiente considerar su característica común de resolver los dilemas operando en el agente un cambio de actitud general. Para un análisis un poco más detallado del tema, ver Apéndice 9.

para el conjunto. Supongamos por ejemplo un juego representado por la siguiente matriz

	B_1	B_2
A_1	(2,2)	(0,3)
A_2	(3,0)	(1,1)

Ejemplo 8

Si este juego se juega de forma no cooperativa, el resultado será (1, 1). El juego puede ser jugado de forma cooperativa aplicando una de las diversas soluciones. Si las soluciones son las numeradas del 1 al 4, veremos que la matriz cambia. Si sucede 1, entonces lo que ocurre es que la única estrategia que los agentes pueden realizar es A_1 y B_1 respectivamente. Si se da la solución 2, la elección de las estrategias no cooperativas A_2 y B_2 queda gravada con algún tipo de penalidad que hace que esta elección deje de ser beneficiosa para el agente (o bien, que la utilización de la estrategia cooperativa esta premiada de un modo que hace que esta elección sea más beneficiosa). En la soluciones 3 y 4, la elección de la estrategia cooperativa también pierde su ventaja, ya que los agentes tendrían unas características morales que les harían atribuir una gran disutilidad a la elección de tales estrategias*. Por tanto, del pago recibido por la utilización de la estrategia no cooperativa habría que descontar esta disutilidad. En el ejemplo 8, es suficiente suponer que esta disutilidad supone una reducción de 2 unidades de utilidad en el

* Debe tenerse en cuenta que el hablar de disutilidades no supone la introducción de ningún concepto problemático. Intuitivamente es fácil de comprender. Que un helado sea de chocolate me hace preferirlo a otros. Ser de chocolate puede medirse positivamente. Es una utilidad. Que al mismo tiempo sea de vainilla le resta atractivo. Ser de vainilla puede medirse como algo negativo. Es una disutilidad.

Una disutilidad no es un concepto extraño. No es un concepto negativo en un sentido extraordinario o estéril. Es una medida cuyo funcionamiento es igual al de la utilidad. En realidad, son la misma cosa. Simplemente, llamamos disutilidad a algo que quita valor, es decir, a una resta en la medida.

En todo caso, el ejemplo no requiere hablar de disutilidades. Podemos suponer que lo que sucede es que el hecho de utilizar una estrategia cooperativa añade una utilidad determinada al valor del resultado. En el caso de las soluciones políticas, concretamente en el caso de la solución 2, utilidades y disutilidades tienen un paralelo en premios y castigos. Un agente puede hacer A o bien porque hacer A se premia o bien porque hacer B se castiga. En este sentido, un premio añade utilidad y un castigo la resta. Un castigo es una disutilidad.

pago definitivo⁷. Entonces la matriz real sería

	B ₁	B ₂
A ₁	(2, 2)	(0, 1)
A ₂	(1, 0)	(1, 1)

Ejemplo 8'

En este caso, ninguno de los dos tiene nada que ganar apartándose de la estrategia cooperativa. La solución cooperativa (2,2) es un punto de equilibrio fuerte, de modo que aun si el juego no es imponible, dos jugadores racionales alcanzarán sin dificultad esta solución. Ambos harán lo que es mejor para ellos y esto, en efecto, será lo mejor. El dilema ha desaparecido.

Lo que sucede en estos casos puede expresarse del siguiente modo. Algunas situaciones parecen un dilema, pero en realidad no lo son. La matriz real del juego no tiene la estructura de un dilema.

Las soluciones que cambian la matriz del juego, transformando lo que originariamente era un dilema en un juego con otras características, sólo son "soluciones" en un sentido. Todas ellas resuelven el problema práctico, haciendo que la elección racional, en el sentido de elección maximizadora de la utilidad individual, sea realmente la elección que colectivamente tiene mejores resultados. Pero no solucionan lo que llamaremos el problema teórico, sencillamente porque este desaparece al desaparecer el dilema. Podemos decir por tanto que estas soluciones en realidad *disuelven* el dilema.

El problema teórico surge cuando un juego es un auténtico dilema irreducible a un juego distinto. La solución en estos casos consiste en

⁷ Puede pensarse que un agente en el que se haya operado una de las transformaciones morales mencionadas siempre asignará una disutilidad tal a la utilización de la estrategia no cooperativa que el resultado de esta será siempre para él menos ventajoso que el resultado de la estrategia cooperativa. Esto no es así necesariamente para ninguno de los cambios morales. Supongamos que se ha convertido en una persona digna de confianza. Puede que esto de lugar a alguna de las soluciones morales. Pero puede que no. Como ya dijimos, puede que este dispuesto a hacer A porque, al ser yo una persona digna de confianza, hacer A ya no es peor para él. Y si este es el motivo por el que yo hago A, e.d., si se da la solución 3, puede ocurrir que en alguna ocasión el hecho de que yo sea una persona digna de confianza no solucione el dilema. Porque puede ocurrir que a pesar de todo A siga siendo peor para él, en cuyo caso no hará A. No romperé ni palabra por un millón de pesetas, pero quizá lo haga por 100. En el caso de la solución 3, todos tenemos un precio.

la elección de la estrategia cooperativa a pesar de que esta elección sigue siendo peor para el agente. Si todos hacemos esa elección, el resultado será mejor para todos. Pero para ello cada uno tiene que realizar una elección que es la peor para él. La solución 5 resuelve, igual que las demás, el problema práctico. Pero, también igual que las demás, deja sin solución el problema teórico. Sólo que, a diferencia de las otras soluciones, la solución 5 no hace desaparecer este problema. Más bien, es tanto que el dilema sigue siendo un auténtico dilema, el problema teórico continúa existiendo.

En esta tercera parte nos ocuparemos del problema teórico. Este problema surge a partir de una determinada teoría sobre la racionalidad práctica, a saber, la teoría que identifica la acción racional con la acción que maximiza la utilidad del agente. Por tanto, parece que es preciso revisar esta teoría. Sin embargo, conviene insistir en algo que ya apunté anteriormente. A pesar de que este problema surja de esta teoría eso no significa que la teoría sea en ningún sentido autocontradictoria. Según esta teoría, un agente racional, enfrentado con un auténtico dilema elegirá la estrategia no cooperativa.

Un juego no cooperativo puede transformarse en un juego cooperativo aplicando las soluciones 1 a 4. Una vez aplicadas estas soluciones, la solución cooperativa será o bien auto-imponible (si se dan las soluciones 3 o 4) o imponible (si se dan las soluciones 1 o 2). En estos casos el agente racional seguirá la estrategia cooperativa. Pero en otro caso el juego será un juego no cooperativo y el agente racional elegirá la estrategia no cooperativa. Dicho de otro modo, unos agentes racionales nunca darán lugar a la solución 5. Elegir la estrategia cooperativa en un juego no cooperativo, e.d., cuando esta elección es peor para el agente es una conducta irracional.

Hasta aquí la teoría. Y no hay nada contradictorio en esto. Precisamente en este punto surge el dilema, porque si todos somos agentes racionales y elegimos la no cooperación el resultado será subóptimo. La apariencia de contradicción se origina precisamente aquí, porque unos agentes racionales según la teoría nunca alcanzarán un resultado óptimo en estas circunstancias mientras que, por otro lado, la teoría exige que un agente racional consiga el máximo de utilidad posible. Y parece que al menos en un sentido, la consecución de la solución cooperativa es pr

sible, a saber, en el sentido de que hay un conjunto de estrategias, una para cada agente, cuya realización daría lugar al resultado cooperativo, y que son una alternativa real en el sentido de que son estrategias disponibles. Pero hay otro sentido importante en el que no es posible alcanzar el resultado cooperativo. En el sentido de que cada agente individual estaría obrando contra sus propios intereses si actuara de forma cooperativa. Y por eso la teoría no es contradictoria. Si cada agente obra según la estrategia que le es más beneficiosa, conseguirá los mejores resultados posibles tal y como lo exige la teoría. Cada uno sale ganado actuando de forma auto-interesada. El problema es que pierde más debido a la actuación auto-interesada de los otros. Pero ¡que le vamos a hacer!. La teoría dice lo que cada uno debe hacer en cada circunstancia. Simplemente, hay circunstancias que imposibilitan que el resultado sea óptimo. Pero la teoría sigue funcionando en estos casos. Dice a cada agente qué es lo mejor que puede hacer y si el agente sigue la teoría, realmente hace lo mejor que puede hacer. Cuando se produce una situación de este tipo, la teoría dice que unos agentes racionales harían bien en transformar el juego no cooperativo en un juego cooperativo. Esto es una tarea política. Pero las soluciones políticas no siempre son aplicables⁹.

Ahora bien, si la teoría no es auto-contradictoria, queda por saber cuál es el problema. Es decir, ¿cómo es posible que cuando cada uno hace la mejor el resultado sea peor para todos?, ¿cómo es posible que unos agentes que mantengan un comportamiento irracional según la teoría

⁹ No resulta tan claro que las soluciones psicológicas-morales 3 y 4 sean, en este sentido, algo que la teoría pueda promocionar, precisamente porque consisten en una transformación del orden de preferencias. La teoría no puede decir: "si eres una persona de confianza y valoras este carácter tuyo lo suficiente como para que la elección de la estrategia cooperativa sea mejor para ti, el resultado (supuesto que los demás tienen también esta cualidad) será óptimo. De modo que procura transformarte en una persona de confianza". Decir esto sería como decir: "si deseas conseguir tu parte del acuerdo todo irá mal. De modo que procura perder tus deseos, aniquila tu voluntad y entonces el resultado será óptimo." En primer lugar, porque si lo que hace óptimo el resultado es conseguir lo que ya no se quiere conseguir, el argumento es falaz. Y en segundo lugar porque si el argumento equivale a decir que no tener deseos es lo mejor para que los deseos no se incumplan, la teoría pierde interés. "¿Quieres conseguir maximizar tu utilidad? Hazla igual a cero", "Sólo hay manzanas, ¿quieres comer lo que deseas? Desea comer manzanas". La argumentación no sería falaz, sería trivial.

Quando el agente tiene esas disposiciones y la elección de A es mejor para él, entonces no hay dilema. Pero cuando lo hay, resulta problemático si la teoría puede proponer la disolución del dilema como salida. Por ello, es mejor referirnos ahora sólo a las soluciones políticas y dejar el análisis de este problema para más adelante (Cuarta parte).

obtengan mejores resultados que unos agentes racionales?

Parfit ofrece una visión sugerente del problema^{*}. Lo que ocurre en estos casos entre un conjunto de agentes racionales es que, si bien cada uno de ellos hace lo que es mejor para él, ellos en conjunto hacen lo que es peor. Es decir, si cada uno de nosotros elige β en una situación de dilema, cada uno está haciendo lo mejor en términos de la teoría, pero nosotros estamos haciendo lo peor. Y a la inversa, si en esa situación cada uno de nosotros hace α , cada uno está haciendo la peor elección, pero nosotros estamos haciendo la mejor.

Por ello, a pesar de que la teoría no es individualmente auto-frustrante puede decirse que lo es colectivamente. Pero esto no quiere decir nada en contra de la teoría, desde el momento en que esta es una teoría acerca de la racionalidad individual y no acerca de la racionalidad colectiva. Por tanto, que la teoría sea auto-frustrante a nivel colectivo no supone una objeción a la teoría¹⁰.

^{*} Parfit, (1984) pp.88-92

¹⁰ Parfit nos previene contra un mal entendido que puede surgir en este punto. En efecto, podría pensarse que la teoría es una mala teoría puesto que no puede aplicarse con éxito a todos los individuos. Pero esto es falso. Nuestra teoría no sólo resulta satisfactoria para un individuo, sino para todos tomados individualmente. En este sentido, la aplicación de la teoría es universal, o.d., se aplica a todo el mundo. Pero no es una teoría colectiva.

CAPITULO VIII LA DERIVACION DE LA MORAL A PARTIR DEL INTERES PROPIO

Las soluciones 4 y 5 presentadas en la introducción de la tercera parte plantean situaciones en las que la moral no tiene por qué coincidir con el interés propio. En concreto, la solución 5 plantea un tipo de situación en la que moral y el interés entran en conflicto.

Esto contrasta directamente con determinadas teorías que intentan derivar la moralidad del interés propio. En capítulos anteriores hemos visto una de estas teorías, la defendida por Gauthier.

Gauthier parte de suponer unos agentes racionales y mutuamente desinteresados, e intenta introducir la moralidad como un medio para la maximización de la utilidad individual. Su hipótesis por tanto es que no sólo no hay conflicto entre moralidad e interés, sino que el mantenimiento de la primera sirve al segundo. Esta hipótesis queda perfectamente expresada al comienzo de su obra principal "El deber predomina sobre el provecho, pero la aceptación del deber es realmente provechosa"¹.

Su intento se desarrolla en tres fases.

1. La primera aparición de la moralidad es la formulación de la salvaguarda lockeana como fuente de derechos en la posición inicial de regateo. El respeto de la salvaguarda se contempla como una moralización del estado de naturaleza anterior a la creación de acuerdo².

Sin embargo, como ya vimos, este primer paso no resulta convincente. Recordemos los motivos. En primer lugar, el único modo sensato de entender la salvaguarda (o cualquier otro conjunto de condiciones de negociación) es como una restricción a lo que cada una de las partes puede llevar a la mesa de negociación como punto inicial de regateo. En este sentido, las condiciones de negociación tienen un papel puramente hipotético y no tienen por qué ser respetadas en la situación real anterior a

¹ Gauthier (86) p.2

² Como ya vimos detenidamente, la razón de introducir estos derechos es que sólo así son posibles acuerdos satisfactorios que son para Gauthier, por ese motivo, estables. Por consiguiente, el comportamiento racional, en tanto que implica el deseo de acuerdos estables, implica la aceptación de unas restricciones morales sobre la conducta individual.

la negociación. En *segundo* lugar, y más importante, no siempre es deseable un acuerdo basado en una situación inicial de regateo que respete las condiciones de negociación. Más bien, las partes beneficiadas por un acuerdo insatisfactorio sólo tienen motivos para desear un acuerdo satisfactorio si el resultado de este es mejor para ellos. Y esto puede no ser así en algunos casos. Y, en *tercer* lugar, un acuerdo satisfactorio no es por ello estable, de modo que el deseo de lograr estabilidad no supone un motivo para la aceptación de tales condiciones.

2. El segundo paso dado consiste en intentar encontrar un motivo auto-interesado para respetar los acuerdos cooperativos. Un agente racional respetará los acuerdos (al menos los satisfactorios) ya que de otro modo será excluido de acuerdos posteriores. Gauthier no niega que sea beneficioso utilizar la estrategia no cooperativa. Lo que afirma es que, a pesar de la ventaja que pueda suponer esto en un caso aislado, a largo plazo esta conducta resulta ser poco rentable, ya que un individuo que haya roto un acuerdo será castigado por los demás con la desconfianza y la consecuente exclusión de acuerdos cooperativos. De modo que el tal individuo acabará perdiendo más de lo que gana.

Como vimos, esta argumentación, si bien no puede ser criticada en sí, no logra los objetivos deseados. En efecto, un individuo que incumple un trato será excluido de otros acuerdos, lo cual sin duda le hará perder. Pero esto sólo sucederá a) si el individuo en cuestión es descubierto, cosa altamente improbable en numerosos dilemas *n-personales*, b) si es posible aplicar la sanción, *e.d.*, si hay una probabilidad suficientemente alta de que los mismos individuos vuelvan a interactuar y c) si el número de interacciones entre los mismos individuos es indeterminado y desconocido por todos los participantes. Con esto no negamos que la cooperación, entendida como conducta cooperativa, no pueda surgir y hacerse estable en algunas circunstancias a pesar de que los acuerdos no sean imponibles³. Lo que negamos, es que esto suceda siempre. Más bien, sucede en muy raras ocasiones. Por ello esto no es suficiente para fundamentar la conducta moral en el interés propio.

3. El tercer paso dado por Gauthier presenta algunas novedades

³ Ejemplos de este surgimiento espontáneo de la cooperación y un estudio detenido de las condiciones que posibilitan este surgimiento pueden encontrarse en Axelrod (86).

respecto a los anteriores. Puesto que en hasta ahora no hemos examinado este paso, le dedicaremos ahora mayor atención. Gauthier es consciente de que si se considera la moralidad como un mero instrumento para el fin del interés propio, entonces se considerará como un mal necesario, como un constreñimiento de la búsqueda del propio beneficio que hay que aceptar para evitar males mayores. Pero si esto fuera así, entonces la conducta moral sólo resultaría atractiva para aquellos que no pueden imponer sus propios términos y para los que saben que serán descubiertos y castigados si actúan de modo inmoral. Para evitar esta consecuencia poco deseable, Gauthier propone una nueva visión del ser humano, distinta de la presentada en la figura del hombre económico y, según él, más cercana a la realidad.

"Para un ser asocial que busca y lucha, la moralidad no puede ser más que un constreñimiento necesario pero que no es bien recibido. Pero para los que valoran la participación la moralidad del acuerdo, a pesar de ser un fuente de constreñimiento, hace que su actividad compartida sea mutuamente bienvenida y, por tanto, estable, asegurando de este modo la ausencia de coerción y engaño".⁴

El texto parece sugerir que el individuo liberal (nombre que se da a este tipo de individuo como figura contrapuesta al hombre económico) valora la participación por sí misma, lo cual hace que su función de utilidad se altere de tal modo que la conducta no cooperativa deja de resultar ventajosa. Si esto es así, entonces tenemos una situación en la que los dilemas se resuelven del modo indicado por la solución 3, e.d., un caso en el que los agentes eligen la estrategia cooperativa. A porque, debido a un cierto cambio de carácter moral en los agentes, A ya no es peor para ellos. Incluso puede que la situación quede mejor definida si eliminamos el término "moral" de la descripción. En realidad, todo lo que sucede es que los agentes, además de querer conseguir determinadas cosas, desean cooperar, de tal modo que la matriz que representa el juego no tiene en ningún momento la estructura de un dilema.

Veamos paso a paso como se llega a esta afirmación:

A) Parte de la argumentación de Gauthier consiste en una afirmación de hecho, a saber, la afirmación de que los seres humanos en muchas ocasiones encuentran satisfacción en la cooperación.

⁴ Gauthier 1981 p.337. La cursiva es mía.

"En muchos y diversos casos, desde las actividades deportivas a las musicales, desde las políticas a las militares, los seres humanos encuentran satisfacción en la participación. Desde luego, la participación debe estar relacionada con algún fin, pero sería erróneo pensar que la participación tiene un valor meramente instrumental. De hecho, alguien puede valorar los fines en parte porque proporcionan ocasión para la participación."²

Según se desprende de este texto, la valoración de la participación no tiene en ningún sentido un carácter moral. Ahora bien, la cuestión fundamental es si esta valoración de la participación, que sin duda se da, es una solución a las situaciones de dilema. Para saber esto, antes es preciso saber si la participación tiene valor como medio o como fin.

Es necesario distinguir aquí dos cuestiones distintas: a) si la participación es siempre participación para algo y b) si la participación se valora como fin o como medio. Según se afirma en el texto, la contestación a la primera pregunta es afirmativa. Sin duda lo es. Uno participa con los demás en un equipo de fútbol para jugar al fútbol, una bailarina colabora con un bailarín para bailar un paso a dos. No se puede participar y no hacer nada participando. Lo que se hace a través de la participación es su producto. En ocasiones podemos hablar de este producto como del fin de la participación, siempre y cuando no asociemos la palabra "fin" con algo que es valioso por sí mismo.

Esto nos lleva a la segunda cuestión. En ocasiones, la cooperación sólo tiene un valor como medio, e.d., sólo es valiosa en tanto que sirve como medio para conseguir un fin. Pero en otras ocasiones, no sólo la participación es el fin, esta vez entendido en el sentido de ser lo valioso por sí mismo, sino que el producto de la participación es casi como una excusa, algo que sólo tiene valor en tanto que da lugar a la participación y, por tanto, algo que tiene valor de medio. El mismo caso de tarea colectiva puede servir para ejemplificar ambos tipos de situaciones. Así, la participación con otro en un partido de frontón puede ser valorada como un medio para lo que aparece como valioso en sí mismo: jugar un partido. Pero también puede ser que jugar el partido sólo sea una ocasión para conseguir algo que se valora por sí mismo: participar con otra persona en algo. Lo importante es que quede claro que las cuestiones a) y b)

² Southier (196), p.325.

son distintas e independientes y que no debe cometerse el error de confundir la evidente respuesta afirmativa a la primera como una respuesta afirmativa a la segunda. Una cosa es que lógicamente el hecho de participar requiera un complemento del tipo para qué o en qué y otra distinta es cuál de las dos cosas consideramos que tiene valor en sí misma y cuál tiene un valor derivado, o, dicho con la terminología habitual, qué tiene valor de medio y qué tiene valor de fin.

Hasta el momento, lo único que ha argumentado Gauthier es que la participación no siempre se valora como medio. Esto es cierto. Pero esto, junto con la afirmación de que de hecho la gente en ocasiones valora la participación como fin, al igual que cualquier otra afirmación de hecho respecto a las preferencias de la gente, no resuelve nuestro problema. Cuando en una situación de interacción lo valioso es la cooperación misma, el agente seguirá la estrategia cooperativa. Pero entonces ya no hay motivo para llamar a esta estrategia A(ltruista), pues al seguirla el agente no obra de un modo beneficioso para los otros jugadores, sino para él mismo (puede que también, y de modo lateral, beneficie a los demás jugadores, pero este no es el motivo por el que un jugador escoge esa estrategia). Una situación en la que todos ganan con la cooperación y ninguno tiene ningún incentivo para obrar de forma no cooperativa no es una situación de dilema. Desear la cooperación como fin no soluciona el dilema, sencillamente porque no hay ningún dilema que resolver. Estas situaciones son, o bien juegos de intereses idénticos, o bien juegos de intereses mixtos en los que la solución óptima es un punto de equilibrio fuerte.

B) Sin embargo, la argumentación de Gauthier no se reduce a la anterior afirmación de hecho. Más bien lo que intenta es explicar *por qué* un agente racional auto-interesado para quien la cooperación sólo tiene un valor instrumental (cosa que incluso puede muy bien sucederle al individuo liberal) ha de cooperar y ha de encontrar valiosa la cooperación. De nuevo, valorar la cooperación se entiende como una característica moral que debe ser fundada en el interés. Y, desde luego, si se argumentara que la conducta moral se basa en el interés porque tenemos interés en

mantener una conducta moral, el argumento sería vicioso*.

¿Por qué entonces debe un individuo racional valorar la cooperación y la participación en tareas colectivas?. Según Gauthier, una persona valora la cooperación (debemos entender que como medio) sólo si considera que esta es un costo necesario para conseguir algún fin valioso. Ese fin valioso por sí mismo es la maximización de la utilidad. Pero, como hemos visto, un agente maximiza su utilidad si no coopera, puesto que la estrategia no cooperativa es mejor para el agente bajo cualquier supuesto acerca de la conducta de los demás. Sin embargo, la conducta no cooperativa generalizada es peor para todos. Este puede ser el motivo por el que se afirma que la cooperación es necesaria para conseguir un fin valioso. Cada uno de los agentes sabe que, supuesto que todos son agentes racionales, el resultado obtenido será el no cooperativo, e.d., un resultado subóptimo. En este sentido, la cooperación es valiosa. "Por supuesto", pensará cada agente individual, "sobre todo la cooperación de los demás. Si los demás cooperan y yo no, el resultado será para mí el mejor de los posibles. Disfrutaré de los beneficios sin tener que pagar los costos". Pero, de nuevo, cada uno sabe que los demás piensan exactamente lo mismo y que, por tanto, la cooperación, cuyo carácter valioso no se duda, no se conseguirá. ¿Qué puede hacer un agente racional?. Sin duda, puede intentar uno de los dos caminos enunciados en la tabla de soluciones. Gauthier parece apuntar hacia la solución 3, e.d., hacia el desarrollo de actitudes morales que hagan que la elección cooperativa deje de ser peor para el agente, y más concretamente, hacia el cambio moral mediante el cual un agente pasa a ser una persona suficientemente altruista. Es más, parece afirmar que dado lo valioso de la cooperación esta actitud se desarrollará espontáneamente. Si los demás se muestran dispuestos a cooperar, yo valoraré su actitud y a partir de esto empezaré a desarrollar un interés por los demás, tanto por su bienestar como por su compañía y por la participación con ellos en tareas conjuntas.

C) A este punto se llega partiendo de la anterior afirmación de hecho mediante un paso intermedio.

* Debe notarse que la solución 3 (ver p.136) no da lugar necesariamente a un argumento circular. Lo único que se dice es que puede suceder que valoremos la conducta moral por sí misma de tal modo que al obrar moralmente obramos en nuestro propio interés. Cuando se da esta situación, moral e interés coinciden.

"Ciertamente las actividades que tienen un valor instrumental deben tener un sentido más allá de ellas mismas, y este sentido no puede ser meramente el compromiso con estas actividades. Debemos por tanto reconocer ciertas estados de cosas intrínsecamente valiosos, deseados por sí mismos. Pero no tenemos que suponer que esos estados de cosas, por muy valiosos que sean, son suficientes para llenar una vida que los seres humanos encuentren digna de ser vivida."⁷

Este argumento no muestra que un agente racional que sólo se interesa en la cooperación como medio deba cooperar. Más bien la argumentación está dirigida a mostrar que, puesto que la vida humana es más valiosa si los hombres cooperan, un individuo racional debe valorar la cooperación. Hay un sentido en el que no puede dudarse de la verdad de esta afirmación. Si se valora la cooperación, la vida tiene más valor al menos en un sentido. Si valoramos la cooperación y cooperamos, la satisfacción que la vida puede proporcionarnos aumenta. Esta argumentación es similar a la que a veces se les hace a las personas a las cuales no les gusta el vino: "No sabes lo que te pierdes. Si te gustara el vino, esto te daría ocasión de aumentar la cantidad de satisfacción en tu vida. Disfrutarías, además de los placeres de los que disfrutas ahora, del placer del vino". Esto probablemente es cierto, al menos si el gusto por el vino es lo suficiente para disfrutar de él cuando llega la ocasión y no tan fuerte como para padecer por su ausencia. Tomado en este sentido, la exhortación a que gustemos del vino, la cooperación o la ópera, es un buen consejo.

Sin embargo, hay otro sentido en el que esto es falso. Puede darse una lectura del texto según la cual se afirma que debe valorarse la cooperación en tanto que es indispensable para conseguir algún fin valioso. El problema está en qué debemos entender aquí por "indispensable". En efecto, con esta expresión podemos referirnos a dos cosas distintas:

A Piénsese en el caso de los prisioneros que da nombre al dilema. Entre los distintos resultados óptimos posibles, hay uno que sería el acordado en un proceso imparcial de regateo. Este resultado, valioso en sí, sólo puede ser conseguido mediante la cooperación. Pero la cooperación no tiene nada que ver con ese resultado. De hecho, el mismo resultado *podría* conseguirse de otro modo (nunca se sabe lo que puede hacer un buen abogado). El resultado en sí es independiente de la cooperación. La

⁷ Gauthier (196), p.352

conexión que une el medio (la cooperación) al fin (el resultado óptimo sobre el que hay acuerdo) no es de necesidad. En otros casos esta conexión es de lo que habitualmente se conoce como "necesidad física"². La conexión de la cooperación con la consecución del fin parece en estos casos bastante sólida. De hecho lo es. Pero ambas cosas, medio y fin, siguen siendo lógicamente independientes.

B Hay, sin embargo, otro tipo de casos en los que esta conexión es de necesidad lógica o necesidad en sentido fuerte. Por ejemplo, supongamos que yo quiero bailar un paso a dos. Bailar un paso a dos *significa* bailar con otra persona. Es imposible que una sola persona baile un paso a dos. Esta imposibilidad no viene sólo dada por el hecho de que, con la ayuda de otra persona, yo pueda hacer cosas que no puedo hacer sola, como por ejemplo hacer 7 pirouettes seguidas o tocar el suelo sólo con la punta de una mano. Quizá alguien pueda dar tantas pirouettes o alguien algún día puede que lo haga. Y sólo debido a la gravedad no puedo yo sin sujeción tocar el suelo sólo con la punta de la mano. Pero estos resultados deseados, igual que en los casos anteriores, sólo de un modo que podríamos llamar accidental tienen algo que ver con la cooperación con otra persona. La imposibilidad de bailar sólo un paso a dos es una *imposibilidad lógica*. Del mismo modo, tampoco puedo jugar un partido de fútbol sin la cooperación con otros. Es decir, hay determinado tipo de actividades que *requieren la cooperación de un modo esencial*.

Quando la conexión entre la cooperación y el resultado deseado es sólo accidental, podemos describirla como una *relación de medios a fines*. En estos casos, la cooperación es sólo un medio. Puede que, por el motivo que sea, ese medio sea de hecho necesario, pero al ser necesario sólo en tanto que medio, la necesidad de su conexión con el fin no elimina la independencia lógica de ambos. Sin embargo, cuando la conexión entre la cooperación y el fin deseado es esencial, hablar de medios y fines pierde su sentido. Si yo quiero bailar un paso a dos, entonces tengo que bailar

²Esto sucede, por ejemplo, en el siguiente caso. Imaginemos un grupo de personas que quiere coger algo escondido en un pozo a 25 metros de profundidad. Si todas colaboran en la tarea, el botín debe ser repartido. Cada una preferiría quedarse con todo, pero el problema es que el tesoro sólo puede ser conseguido mediante la cooperación, e.d., si forman una cadena agarrándose unos a otros (supongamos para simplificar que la situación hace imposible el uso de escalas etc). Uno sólo podría conseguirlo si algo cambiara en la naturaleza, por ejemplo si uno de ellos midiera al menos 24 metros, si pudiera dar un salto vertical de esa longitud o si pudiera volar.

con otra persona. Pero bailar con otra persona *no es* un medio para bailar un paso a dos. Más bien, bailar con otra persona (ateniéndose a determinadas reglas) *es* bailar un paso a dos. La relación entre ambas cosas no es causal, sino formal. En realidad, esto es lo que quiere decir que la relación entre dos cosas es de necesidad lógica.

Recordemos que lo que nos interesa es saber si tiene razón Gauthier al afirmar que un agente racional valorará en sí la cooperación cuando esta es un medio para conseguir algún fin valioso por sí mismo. La distinción entre estos dos tipos de casos es sumamente importante para resolver nuestra cuestión⁹. Porque la relación de preferencia se establece de un modo muy distinto en uno y en otro. Kant afirmaba que el que quiere un fin quiere también los medios indispensables para alcanzarlo¹⁰. Kant se refiere a aquellos medios cuya relación con el fin se expresa mediante proposiciones sintéticas.

"Que para dividir una línea en dos parte iguales, según un principio seguro, tengo que trazar desde sus extremos dos arcos de círculo, es cosa que la matemática enseña, sin duda por proposiciones sintéticas; pero una vez que sé que sólo mediante esa acción puede producirse el citado efecto, si quiero integro el efecto, quiero también la acción que es necesaria para él, y esto último sí que es una proposición analítica, pues es lo mismo representarme algo como efecto posible de cierta manera por mí y representarme a mí mismo como obrando de esa manera con respecto al tal efecto."¹¹

Tales proposiciones sintéticas expresan auténticas relaciones causales. Es más, ninguna relación auténticamente causal puede expresarse mediante proposiciones analíticas. El primer problema que surge respecto a estas relaciones causales es que es sumamente difícil decir si son o no necesarias y en que sentido. Esto sucede porque, en tanto que no haya ninguna conexión esencial entre medio y fin, siempre existe la posibilidad de obtener el fin utilizando otro medio. Sea como sea, hay muchos

⁹ No debe confundirse esta distinción con la establecida más arriba (p.148). No se trata ahora de que la cooperación pueda ser en ocasiones un fin en sí misma. Más bien se trata de que hay casos en los que no hay una relación causal de medios a fines entre la cooperación y algo que sea resultado de la misma ni en una dirección ni en otra.

¹⁰ Kant (81) p.66.

¹¹ Kant (81) pp. 66-7.

sentido propio) es necesaria de hecho, e.d., en los que la consecución del fin en un momento determinado pasa imprescindiblemente por la utilización de determinados medios. Supongamos que esto es suficiente para decir que un medio es "indispensablemente necesario para alcanzar un fin".

¿En qué sentido puede decirse que, en estos casos, el que quiere el fin quiere (o, mejor, debe querer) los medios? Hay un sentido trivial en el que esto es verdad, a saber, en el sentido de que el que quiere un fin tiene que estar dispuesto a poner los medios. Pero hay otro sentido en el que esto no es cierto. En este otro sentido, no sólo el que quiere el fin no tiene por qué querer los medios, sino que incluso, y en muchas ocasiones, quiere el fin a pesar de los medios que es necesario disponer, e.d., a pesar de no querer los medios. Esto es precisamente lo que diferencia querer algo como fin y quererlo como medio. Lo que se quiere como medio se quiere en el sentido de que se está dispuesto a realizarlo, pero no en el sentido de que su utilización sea, digamos, deseada. En muchos casos el empleo de determinados medios supone un coste a descontar del total de utilidad que supone la consecución del fin. De hecho, cuando hay más de un medio disponible, se escoge aquel que resulta menos costoso. Cuando sólo hay uno, no existe esta posibilidad de elección, pero el medio puede seguir resultando costoso. Es más, esto sucede en todos los casos en los que el medio tiene sólo valor como tal. Como mucho, en algún caso puede no resultar costoso, sin que esto suponga que tiene un valor positivo.

En los casos de tipo B, en los que la relación no es causal sino formal ocurre otra cosa muy distinta. Pero no porque el que quiera una cosa deba querer la otra, sino porque el que quiere una cosa, *de hecho*, está queriendo la otra. Si yo quiero bailar un paso a dos entonces quiero bailar con otra persona. Si quiero jugar un partido de fútbol, entonces quiero jugar con otras personas. Pero aquí no hay fines ni medios en sentido propio. No hay relación causal.

Sin duda, y a juzgar por los ejemplos utilizados por Gauthier, cuando este afirma que, al querer un fin valioso por sí mismo, un agente racional querrá también los medios, está pensando en este tipo de casos. Uno de sus ejemplos más utilizados es el de la orquesta, donde afirma que "incluso si un sólo individuo pudiera llegar a dominar todos los instru-

mentos, no podría convertirse en una orquesta"¹². Gauthier habla como si en estos casos la cooperación fuera un medio necesario. Pero no lo es. No es un medio de ningún tipo. Más bien el fin que se quiere es un fin, digamos, colectivo. Yo quiero tocar en una orquesta. Esto mismo ya es querer participar con otros¹³. Pero no es que yo valore en sí un medio.

Gauthier afirma que el individuo liberal, en estos casos valorará el medio, la participación. Esto es una de las características que le distinguen de otro tipo de hombre, el hombre económico. Pero, por todo lo dicho, esto no es así. Naturalmente que el individuo liberal valora la participación en estos casos, pero esto no le distingue de los demás. Por el contrario, todo aquel que se proponga como fin uno de estos "fines colectivos" lo hace. Porque en estos casos, la afirmación no es, como él pretende, una afirmación de carácter psicológico. La psicología no tiene nada que ver aquí. Es una afirmación de tipo lógico, en la que se muestra cuál es la implicación lógica de querer un fin colectivo, a saber, el querer lo que, o bien forma parte de este fin, o bien es el fin mismo.

D) El siguiente paso en la argumentación consiste en afirmar que, no sólo cuando un individuo quiere uno de estos fines quiere por ello la participación, sino que estos fines son tales que *deben* ser queridos.

"¿No podemos más bien suponer que nuestra capacidad social para encontrar valor en la participación es una de las principales fuentes de enriquecimiento en la vida humana, haciendo posible, como lo hace, la realización complementaria de nuestras variadas capacidades y potencias humanas?"¹⁴.

Naturalmente, la contestación de Gauthier es afirmativa. Es importante notar que el planteamiento de la cuestión retrotrae la argumentación al punto que discutimos más arriba (p.150), en el momento en el que se planteaba si debe valorarse la cooperación. La respuesta, en este ca-

¹² Gauthier (86) p.336.

¹³ Es notable que, al hablar de estos casos, Gauthier utilice indistintamente los términos "cooperación" y "participación". Sin embargo, es importante notar que no son equivalentes. Tanto los casos que plantea Gauthier como los que yo presento son casos de participación. Pero esto no excluye la posibilidad de que se planteen situaciones de intereses mixtos en los que la estrategia no-cooperativa sea la única que da lugar a un punto de equilibrio. Puedo querer jugar una partida de póquer. Esto significa que quiero participar. Pero también puedo utilizar con los demás una estrategia no cooperativa, como hacer trampa.

¹⁴ Gauthier (86) p.337.

so, vuelve a ser la misma: al igual que adquirir gusto por el vino o la música clásica o cualquier otra ampliación del repertorio de las cosas capaces de producirnos satisfacción, adquirir gusto por las tareas colectivas incrementa nuestras posibilidades de satisfacción. Esto es cierto. Incluso podemos dar por bueno que estas tareas no sólo son una fuente de enriquecimiento, sino que es una de las principales, siempre y cuando se entienda correctamente. Es una de las fundamentales fuentes de enriquecimiento porque es una de las más numerosas y variadas. Pero la cuestión acerca de si las satisfacciones de este tipo son "en sí" más valiosas o no lo son es sin comparación mucho más discutible. De hecho, nuestra teoría afirma que las preferencias (racionales) de los agentes sólo son más o menos valiosa en tanto que estos las colocan en uno u otro lugar en la ordenación. Pero esto hace aun más dudosa la afirmación de que un agente racional *debe* valorar estas actividades, igual de dudosa que la afirmación de que un agente racional *debe* valorar el vino. Esto es cuestión del contenido de las preferencias, y la racionalidad, al menos tal y como la presenta la teoría de la que hemos partido, que es por cierto la misma de la que parte Gauthier, no hace ningún supuesto acerca del contenido de las preferencias. No resulta lícito hacerlo ahora sólo para salir del paso.

Ateniéndonos a los casos en los que la cooperación es realmente un medio, e.d., a los casos A, ¿por qué no considerar la conducta cooperativa como un mal menor, como un medio cuyo necesario empleo puede restar valor al resultado?. Según Gauthier, esto sucede porque el resultado de la cooperación se considera únicamente como preferible en segundo lugar. Pero, añade, esto es falso. De hecho, pretende mostrar que el resultado de la conducta cooperativa es el mejor resultado. El argumento fundamental es que la cooperación nos permite conseguir resultados que, sin ella, estarían fuera del alcance de las posibilidades humanas. Y este hecho no es de lamentar, pues nos permite gozar de la cooperación. Pero aunque es cierto que el resultado de la cooperación no puede ser alcanzado sin esta, como hemos mostrado, que esto no sea lamentable depende de si se desea la cooperación y, en caso contrario, el resultado de la cooperación *sí es lo que prefiero en segundo lugar*. Pensemos en los prisioneros del dilema. Para cada uno hay un resultado mejor que el resultado cooperativo, a saber, aquel resultado que se sigue si el otro no confiesa y uno

sí. Esto sucede en todos los dilemas de este tipo. Naturalmente, hay situaciones que no son de dilema. En ellas el resultado "cooperativo" es el mejor, pero lo que necesitamos es una teoría que solucione los dilemas, no que suponga que estos no se dan.

Es importante señalar que en los casos en los que el resultado cooperativo es considerado como el mejor la elección de la cooperación obedece directamente a motivos de interés. Puede suceder que lo que haga preferible este resultado sea que los agentes tienen interés en los intereses de otros o cualquier otra cosa del tipo que hemos identificado como solución 3. Pero esta solución sólo lo es del problema práctico, mientras que el problema teórico queda disuelto.

E) La argumentación de Gauthier fracasa en demostrar precisamente lo que era su tesis principal, a saber, que cuando la participación se valora como medio, pasa (o, incluso, debe pasar) a valorarse como fin. Cuando la participación esta en relación causal con alguna otra cosa y en esta relación la participación desempeña el papel de medio, entonces no hay ningún motivo para que pase a valorarse como fin. Sin embargo, el argumento no se detiene en este punto. Porque, según él, de la valoración de la participación, e.d., de la valoración que no es meramente instrumental, se pasa a valorar a los que participan con nosotros.

"Pero además de esto, podemos suponer que, al valorar la participación, una persona llega a valorar a sus compañeros, de tal modo que las actividades compartidas dan lugar a vínculos entre las personas *que conducen a cada uno a tomarse interés en los intereses de los otros*- aunque, desde luego, no en el interés de todos los otros sino sólo de aquellos a los que se identifica y experimenta como co-participantes."¹³

Esto sucede porque

"Las personas llegan a tomarse interés en sus compañeros porque reconocen la mutua disposición a no aprovecharse unos de otros y a compartir de un modo justo los beneficios conseguidos mediante una actividad compartida."¹⁴

Esto significa que las personas desarrollarían un altruismo suficiente como para resolver el problema práctico planteado por el Dilema

¹³ Gauthier (86) p.336. La cursiva es mía.

¹⁴ Gauthier (86), p.338

del prisionero. Más concretamente, los dilemas se solucionarían al escoger los jugadores la estrategia A por el motivo 3, e.d., porque, debido a que ahora son lo suficientemente altruistas, A ya no es peor para ellos. Como los intereses de los demás llegan a valorarse a partir de la valoración de la cooperación, y esta es racional, podemos decir que la moral se basa en el interés propio o, dicho de otro modo, podemos afirmar que partiendo del supuesto de unos agentes auto-interesados hemos llegado a derivar la moralidad¹⁷. Lamentablemente, la argumentación no parece concluyente.

Para Gauthier, el hecho de participar en tareas conjuntas con otras personas, junto con el hecho de que estas se muestren dispuestas a basar la cooperación en unos términos justos, es suficiente para que cada uno desarrolle un interés altruista en los intereses de los demás. Aunque hasta el momento todo el peso de la argumentación ha estado en los casos que nosotros hemos llamado de tareas colectivas y en aquellos en los que la cooperación es valorada como fin, en este momento la conclusión se extiende a todos los casos de cooperación. Sin embargo, este paso no es del todo legítimo. Pensemos en el paso a dos. La matriz de este juego podría ser

	B ₁	B ₂
A ₁	(1,1)	(0,0)
A ₂	(0,0)	(0,0)

Ejemplo 9

donde las estrategias A₁ y B₁ son las de cooperar y A₂ y B₂ las de no cooperar. El "objetivo" de la cooperación es bailar un paso a dos. Pero para bailar un paso a dos hay que contar con otra persona, por tanto, puesto que ambos quieren bailar un paso a dos, ambos quieren cooperar. Si los dos cooperan, se baila el paso a dos, y si alguno o ambos no coopera, no se baila (suponemos para simplificar que sólo hay dos bailarines disponibles). Puesto que ambos quieren bailar, la matriz será de este tipo. Pero, como es fácil ver, esta matriz no representa ningún juego de inte-

¹⁷ En este momento, entendamos por "conducta moral" simplemente conducta cooperativa.

reses mixtos. Es la matriz de un juego de intereses idénticos¹⁸. Lo mismo sucede en los casos en los que la propia cooperación es el fin deseado por todos los jugadores¹⁹.

Por consiguiente, en estos casos el dilema no se plantearía. El hecho de que cada jugador se interese o no en el bienestar de los otros no soluciona ningún problema porque no hay ningún problema que resolver. La afirmación de Gauthier es en estos casos superflua. Pero, además, la generalización del caso de las actividades colectivas y del caso en el que la cooperación es el fin a la totalidad de las situaciones en las que se plantea la cooperación es ilegítima, al tomar toda su plausibilidad de los casos en los que la cooperación es el fin. Por una razón. Cuando para alguien el producto de la cooperación es sólo una excusa para la cooperación misma, lo más habitual es que esa persona sienta una cierta simpatía por las personas con las que quiere colaborar. Si preparar unas oposiciones con otra persona es para mí sólo una excusa para hacer algo con esa persona, por lo general lo que ocurre es que yo quiero sentirme unida a ella y esto significa que mi interés por ella no es instrumental. Pero no es que la cooperación haga surgir la simpatía. Es que la existencia de esta es el principal (si no el único) motivo para considerar la cooperación con otra persona como un fin en sí mismo. De modo que la existencia del interés por los demás en estos casos no prueba lo que Gauthier quiere probar, a saber, que es la actividad compartida la que genera la simpatía. Y, en el resto de los casos, incluidos la mayoría de las actividades

¹⁸ Naturalmente, dos personas en esta situación pueden tener algunos intereses contrapuestos, como por ejemplo el puesto en el que aparecerán en el cartel, el tipo de letra en que irá su nombre o como saludarán. Si ponerse de acuerdo en estas cuestiones depende de ellos (cosa que no suele ocurrir) deberán entablar un proceso de regateo. El resultado será garantizado en este caso por la existencia de un contrato. Pero el regateo no tendrá como objeto el determinar si bailan o no ni hasta que punto bailan. Más bien, es probable que el bailar o no sea utilizado por cada uno como una amenaza para conseguir en el regateo el mejor resultado posible. Pero esto no nos interesa ahora, porque la terna colectiva, de la cual nos estamos ocupando, es bailar un paso a dos. La composición del cartel no es una terna colectiva.

¹⁹ En el caso de que algunos jugadores quieran la cooperación como fin y otros como medio esto no sucede, sino que el juego sería de intereses mixtos. Pero tampoco en esos casos se plantearía el dilema (al menos en los juegos bipersonales y en los juegos n-personales en los que un número suficiente valore la cooperación como fin) puesto que la elección de A por parte de los jugadores que valoran la cooperación como fin sería suficiente para asegurar un resultado óptimo.

colectivas²⁰, esto no sucede. Aquí el interés por los demás es puramente instrumental o, al menos, no es necesario que suceda de otra manera.

Gauthier afirma que, al encontrar a los demás dispuestos a la cooperación, se desarrolla un interés por los demás y por la cooperación con ellos. Pero ¿por qué?. Los demás, al igual que yo, están dispuestos a cooperar porque en determinadas circunstancias eso les resulta ventajoso. ¿Por qué esto ha de hacerme tomar interés en sus asuntos?. Si un colaborador esencial e insustituible cae enfermo lo sentiré mucho, pero sólo en tanto que eso afecta a la consecución de los objetivos comunes. A lo mejor me cae bien. Incluso es posible que, gracias a la participación en una tarea común, haya llegado a conocerle mejor y me sea simpático. Sin duda, a los seres humanos a veces les cae bien otro ser humano. Pero a veces no. Mi compañero, incluso cuando es insustituible, puede ser un pelmazo, o tener la fea costumbre de hablar con la boca llena y a gritos. Esto es igualmente cierto en los casos de las tareas esencialmente colectivas. Aunque querer bailar un paso a dos signifique querer bailar con otra persona, eso no me obliga a sentir ninguna simpatía por mi compañero de baile ni a tomarme ningún interés por sus asuntos más allá de que no se rompa un tobillo o se maree antes de la sesión de noche. Y este interés es el que siente un piloto por su coche. De modo que, aun en el supuesto caso de que fuera cierto que las actividades colectivas enriquecen la vida humana de un modo que las hace indispensables, de aquí no se sigue necesariamente que se genere espontáneamente ningún sentimiento altruista. Y respecto a que pase de encontrar la cooperación valiosa como medio a encontrarla valiosa en sí, depende fundamentalmente de lo anterior y no al revés como parece pretender Gauthier. No es que primero pase a valorar la cooperación en sí y después a los que cooperan conmigo. Más bien al contrario, si sucede que la cooperación me da ocasión de conocer a una persona que resulta que me es extraordinariamente simpática, y llego a tener interés no instrumental por sus asuntos, pasará a encontrar en la cooperación un valor no dependiente sólo por el hecho que de me dará

²⁰ Dijo que en la mayoría y no en la totalidad porque puede suceder que en una actividad colectiva lo único que uno busque principalmente sea la, digamos, parte colectiva, siendo más o menos indiferente a lo demás. Estos casos son semejantes a aquellos en los que la cooperación es el fin. Por ejemplo, yo puedo querer bailar con otra persona porque quiero bailar un paso a dos, pero también puede que bailar un paso a dos sea una excusa para bailar con otra persona.

ocasión de seguir estando con esa persona unidos por una tarea común. Pero esto no sucede siempre. Hay gente con la que es un placer en sí participar en tareas colectivas y en todo tipo de actividades cooperativas. Con otros es un tormento.

Gauthier dice que estas afirmaciones son de carácter psicológico. Desde luego, lo son. Sólo que, por las razones expuestas, yo pienso que son dudosas. Más bien tiendo a pensar que en muchos casos es difícil que ocurra de ese modo, especialmente en las situaciones de dilema *n*-personal donde las relaciones de reciprocidad quedan perdidas en el anonimato. Pero en todo caso esta discusión queda fuera del alcance de este trabajo. Entonces, ¿qué es lo que en estas afirmaciones es realmente importante? Lo que se pretende al hablar del surgimiento de los sentimientos altruistas es una "reconstrucción racional". Es decir, el propósito de Gauthier no es mostrar cómo surgen los sentimientos morales, sino mostrar que los sentimientos morales que de hecho tenemos (sea cual sea la forma en la que han surgido) tienen una *justificación racional*. Puesto que Gauthier parte de la teoría acerca de la racionalidad práctica que nosotros estamos utilizando, esto quiere decir que nuestros sentimientos morales se justifican racionalmente en tanto que su mantenimiento es un medio para la maximización de la utilidad o, al menos, si no es contradictorio con ella. Por tanto, el propósito de toda la argumentación es derivar la moralidad del interés propio. Ahora bien, ¿muestran esto los argumentos que hemos expuesto y analizado en este capítulo? Lo que se ha pretendido mostrar es que

"una persona, reflexionando sobre sus sentimientos morales, los consideraría una extensión apropiada de su interés por los otros en el contexto de actividades participativas valiosas, y entonces los consideraría esas actividades valiosas apropiadamente en tanto que están sujetas a constreñimientos morales."²¹

Nosotros hemos visto que la cooperación no tiene por qué generar ningún sentimiento de interés por los asuntos de los demás. Esta afirmación no es psicológica. No dice nada acerca de cómo surgen o dejan de surgir los sentimientos morales. Lo que dice es que no hay nada en las tareas cooperativas, ni siquiera en la necesaria cooperación requerida por las actividades colectivas, que haga suponer que un agente racional

²¹ Gauthier (1981), p. 339.

encontrará que sus sentimientos morales (si los tiene) tienen en ellas su justificación, ni que haga pensar a un agente racional que carezca de tales sentimientos que haría bien en tenerlos.

La última parte del texto citado hace referencia a por qué deben ser valorados los constreñimientos morales. Si se pretende fundamentar la moral en el interés, la razón más directa sería que las constricciones morales, e.d., que el obrar moralmente, promueve el interés propio. Lamentablemente, esto no es así, salvo en el caso trivial (solución 3 del dilema) de que estemos interesados en obrar moralmente. Otra posible razón es que los constreñimientos morales deben ser valorados porque hacen posible la cooperación y esta, a su vez, debe ser valorada en sí misma. Pero la validez de esta respuesta depende del supuesto de que la cooperación es valiosa en sí misma, supuesto que hemos rechazado. El único sentido en que esto puede mantenerse es en el sentido de que un agente racional encontrará más valiosa la conducta cooperativa general que la no cooperación generalizada. Por ello, un agente racional estará dispuesto a cooperar sólo si su cooperación es condición necesaria para garantizar la cooperación de los demás. Pero como ya mostramos anteriormente (Cap.VII, pp.127 y ss.), esto ocurre en contadas ocasiones.

Su argumento es que un agente racional, *al encontrar valiosa la cooperación como medio*, pasará también a encontrarla valiosa en sí. No sólo no hay ningún motivo para sostener esto, sino que, además, los dilemas surgen precisamente en los casos en los ni siquiera se da a la cooperación valor instrumental. Salvo en muy determinadas circunstancias de reciprocidad directa e interacción repetida, un agente racional no encontrará valiosa la cooperación como medio. Y en aquellos casos en los que se valora la cooperación como medio, no parece haber motivos para suponer que se pase a valorar como fin. De modo que, aun aceptando que los constreñimientos morales hacen posible la cooperación, aquellos sólo se valorarán cuando se valore esta, lo que no sucede de modo necesario ni en la mayoría de los casos y, lo que es más importante, no sucede en los casos que presentan un problema, a saber, en los casos de dilema.

Queda por saber, si, en el caso de que Gauthier hubiera demostrado lo que pretende demostrar, esto habría sido una derivación de la moralidad. Si el argumento fuera válido, se hubiera mostrado que un agente racional acaba por valorar la cooperación en sí y los intereses de sus com-

pañeros. A pesar de que ese interés por el prójimo tendría su origen en el interés propio, lo trascendería y se haría autónomo, e.d., por ejemplo, haría que se tuviera interés por los demás incluso cuando la interacción cooperativa hubiera terminado. Es problemático que este interés se extendiera también a personas con las que nunca se coopera de hecho o a los que es imposible incluir en tareas cooperativas. Los habitantes de sociedades lejanas formarían parte del primer grupo. Los que no tienen nada que ofrecer formarían parte del segundo²². Para evitar problemas, supondremos una sociedad en la que todos entablan relaciones cooperativas con todos. En esta sociedad, cada miembro tiene un interés en el bienestar de todos los demás. En ella, las consideraciones de carácter moral comprometen no sólo la razón de los agentes, sino también (sobre todo) sus sentimientos. Esto es una forma de decir que los agentes elegirán siempre la estrategia A puesto que, debido al interés que sienten unos por otros, A no es peor para ellos, es decir, en esta comunidad se da la solución 3 de los dilemas. Gauthier afirma que una situación de este tipo puede tener dos orígenes distintos y que, por tanto, hay dos modos en los que la moralidad puede comprometernos afectivamente²³.

1. Puede suceder que estas personas tengan capacidad para lo que podríamos llamar una "moral afectiva", que no es sino una forma de altruismo universal. Es una moral basada en sentimientos de benevolencia. Pero si concebimos la moral básicamente como un conjunto de constreñimientos en la búsqueda del interés propio, salvo que definamos este de una forma meramente egoísta, la situación queda mejor descrita diciendo que, entre personas con este grado de benevolencia, la moral es innecesaria. Los intereses de cada uno están de tal modo determinados por consideraciones acerca del interés de los demás que, sin necesidad de constreñimientos de ningún tipo, todos elegirían siempre la estrategia A. Es decir, todos harían A porque A sería la estrategia que maximizaría su utilidad. Que los intereses de cada uno sean altruistas no es más que un dato del mismo tipo que el que sean egoístas. La única diferencia es que

²² Una de las objeciones más comunes a la teoría de Gauthier es que deja fuera de la relación moral a todos aquellos con los que no es posible entrar en una relación cooperativa fructífera, tales como determinados inválidos físicos y psíquicos.

²³ Gauthier (86), p.327.

entre este tipo de personas, siempre y cuando su altruismo sea moderado²⁴, no se plantearían dilemas. Y no se plantearían porque no sucede que cada uno tenga un interés distinto del interés de los demás y que pueda entrar en conflicto con ellos. Más bien, todos tienen el mismo interés, a saber, que los intereses de todos se satisfagan de la forma más completa posible. Esto haría la moral innecesaria. Muchos autores han señalado que el supuesto de la benevolencia universal anula las condiciones que hacen necesario y posible el surgimiento de la moral. Por ejemplo, Hume afirma que la moral es necesaria porque a) los recursos con los que se cuenta para satisfacer los intereses son escasos y b) no existe la benevolencia universal²⁵. Hay un sentido en el que la conducta basada en la benevolencia universal es en sí misma una conducta moral a pesar de no ser una conducta constreñida, a saber, en el sentido de que nuestra conducta sería idéntica a la conducta que obedece a reglas morales. Concretamente, la existencia de esta moral afectiva haría que todos los agentes eligieran siempre la estrategia A. Pero esta coincidencia conductual no es suficiente para que podamos hablar de moral. Más bien lo que caracteriza la conducta moral, distinguiéndola de otros tipos de conducta, es precisamente lo que distingue la solución 3 de las soluciones 4 y 5, a saber, el hecho de que la conducta moral representa una constricción. Cuando se da la solución 3 los agentes no se enfrentan a ningún conflicto de elección, no hay ningún conflicto entre A y otra conducta posible a la que llamamos B. Pero tampoco sucede que esta conducta altruista se fundamente en el interés propio. En este caso, es verdad que A es la estrategia justificada por el interés, pero sólo en el sentido trivial de que la conducta A es al mismo tiempo B.

Se podría objetar que lo que caracteriza la conducta moral es que los agentes tengan este sentimiento de benevolencia. Por ejemplo, Parfit sostiene algo parecido al pedir que el interés propio se defina de un modo independiente de las preferencias de los agentes. Es cierto que no se puede hacer que el interés propio sea equivalente por definición a

²⁴ Dos altruistas puros también se enfrentarían a dilemas (ver Cap. IX y Apéndice 9).

²⁵ Sobre las circunstancias que hacen posible el surgimiento de la moral ver Hume (81), pp. 708 y ss.

aquello que los agentes promocionan mediante su conducta. En este sentido, Parfit tiene razón al rechazar afirmaciones del tipo "lo que cada uno hace es, por definición, lo mejor para él"²⁴. Pero eso no significa que el interés propio deba definirse de una forma estrecha, entendiendo que "interés propio" hace referencia únicamente a intereses egoístas. Cuando una madre hace lo que es mejor para su hijo esta, por lo general, obrando interesadamente. Y si suponemos que el interés de los agentes benevolentes por los demás es un interés de este tipo, entonces cuando eligen A están obrando de un modo auto-interesado. Desde este punto de vista, la así llamada "moral afectiva" no se distingue de la conducta interesada. Esto es precisamente lo que se quiere decir al afirmar que la existencia de la benevolencia universal no sólo no es una forma de moral, sino que es una de las condiciones que hacen la moral innecesaria. Por otro lado, no resulta muy satisfactorio identificar la moral con algo que no esta sujeto a la voluntad. Habitualmente entendemos que la moral se expresa en prescripciones que podemos obedecer. Y nadie puede obedecer una prescripción del tipo "ten sentimientos benevolentes".

Dejando aparte que la existencia de la benevolencia universal haga innecesaria la moralidad, está el hecho fundamental de que tal sentimiento no existe, salvo quizá muy excepcionalmente. Esto hace que la solución 3 asociada al altruismo no se de en la práctica.

2. El segundo modo en el que las consideraciones morales pueden comprometernos afectivamente es mediante una *capacidad afectiva para la moral racional*. Tener esta capacidad significa poder sentirse emocionalmente comprometido con aquello que reconocemos como una consideración moral válida. Esta capacidad, contrariamente a lo que sucedía en el caso anterior, presupone una concepción previa de la moralidad puesto que "uno no puede ser movido por un sentido del deber a menos que uno, con anterioridad, crea que alguna acción es su deber"²⁷. Una capacidad de este tipo daría lugar a la solución 4. Los agentes así dispuestos harían A independientemente de los beneficios que A les proporcione, y sólo es una consecuencia posterior el que, al estar así dispuestos, A ya no resulte

²⁴ Parfit (86), p. 87.

²⁷ Gauthier (86), p. 328.

peor para ellos. Para Gauthier, esta actitud y, por tanto, esta solución, es la que propiamente merece el nombre de "moral"

"Entendido en sentido propio, un hombre justo es aquel que, reconociendo un determinado curso de acción como justo, encuentra sus sentimientos comprometidos por ese reconocimiento y así se encuentra a sí mismo dispuesto a adherirse a tal curso de acción a causa de su justicia"²².

No es mi intención contradecir a Gauthier en este sentido. A pesar de que a mi juicio es preciso hacer determinadas matizaciones a esta afirmación, creo que es básicamente correcta. La cuestión es si la moral así caracterizada puede basarse en el interés propio tal y como él pretende. Para poder contestar de modo afirmativo, sería necesario que el agente racional a) encontrara que la constricción moral de la conducta puede ser justificada en términos de interés propio, pues esto es la condición que hace que las constricciones morales sean racionales, y b) que posteriormente se encontrara afectivamente comprometido con tales constricciones. A lo largo de todo este capítulo hemos negado la primera de estas condiciones. Pero la cuestión que debemos contestar ahora es si, en el caso de que la condición a) se cumpliera, esto crearía un compromiso afectivo del agente racional con la moralidad.

Para esto podemos fijarnos en los casos en los que sí ocurre que la conducta cooperativa resulta beneficiosa en términos de interés, por ejemplo en un caso de un dilema que se plantea a los mismos agentes en un número indeterminado de veces. En estos casos, un agente racional elegirá la estrategia A porque, sin necesidad de hacer ningún supuesto sobre el carácter moral de este agente, esta conducta le resulta beneficiosa. Según Gauthier, un individuo liberal pasaría a encontrarse afectivamente ligado a la moral. "Sentirse afectivamente ligado a la moral" significa sentirse inclinado a realizar la conducta dictada por la moral por la razón de que esta es precisamente la conducta dictada por la moral. Pero suponer esta capacidad afectiva para la moral es innecesario, puesto que si la conducta moral beneficia al agente este se sentirá inclinado a ella de todos modos. Sin embargo, esto no hace justicia a Gauthier. La elección de A se justifica racionalmente porque es la conducta más beneficiosa a largo plazo, pero en cada caso concreto de elección, A puede re-

²² Gauthier (86), p.328.

sultar perjudicial. A pesar de esto, un individuo liberal hará A porque, una vez reconocida la justificación de la moral en términos de interés, se siente emocionalmente comprometido con esta conducta. A su vez, este compromiso hace que A tampoco sea perjudicial en el caso concreto. Pero esto no significa que la moral sea trivialmente idéntica al interés propio, puesto que primero exigimos que la conducta moral tenga una justificación basada en el interés.

Aun entendido en este sentido, es posible pensar que el compromiso afectivo sigue siendo inoperante, pues un agente racional elegirá A en esos casos de todos modos porque A es la mejor conducta a seguir. En todo caso, el compromiso afectivo puede servir para garantizar la elección de A en los casos en los que el agente, debido, por ejemplo, a lo prolongado de la interacción, puede "olvidar" que A es beneficiosa. Un agente puede verse tentado a utilizar B por los beneficios inmediatos que esta conducta le reporta a pesar de que esto le perjudique a más largo plazo. Por eso, un compromiso afectivo con la moralidad, que nos disponga a elegir A con independencia del resultado, resulta útil. Pero es dudoso que Gauthier piense en estos casos, ya que los agentes racionales ideales en los que basa su teoría no tienen "olvidos" ni "tentaciones" de este tipo.

El compromiso afectivo no sólo resulta inoperante sino que presenta cierta inconsistencia con la propia teoría de Gauthier. Si un agente sólo encuentra justificada la conducta moral cuando esta promueve el interés, no se entiende bien qué quiere decir que, una vez que ha visto la moral así justificada, está dispuesto a adoptar esta conducta con independencia de los resultados que esto tenga para él en términos de interés. Según Gauthier, el individuo liberal "no busca aprovecharse ni mejorar su situación con respecto a los demás procurando violentar unos restricciones justificadas sin apelar al mutuo concernimiento."²⁷

Pero si los restringimientos morales están justificados en términos de interés, (que no lo están) malamente pueden ser violentados en el propio provecho, ya que por hipótesis lo que está en favor del propio interés es mantenerlos.

En resumen, Gauthier intenta dar una justificación racional, e.d., auto-interesada, a la moralidad. Esta fundamentación sólo puede hacerse

²⁷ Gauthier (86), p. 329.

de dos maneras. O bien se entiende que estamos interesados en la moral (solución 3) cosa que Gauthier descarta, o bien mostramos que obrar moralmente promueve nuestro interés, definido este de modo independiente. Esto es lo que Gauthier intenta. Pero, con las escasas excepciones mencionadas, esto no sucede. Simplemente, no es cierto que la virtud siempre (ni siquiera casi siempre) obtenga recompensa, y el caso en el que la virtud es su propia recompensa queda descartado por irrelevante. Pero si fuera cierto, lo que se mostraría es que no hay dilemas, no que haya una solución moral a los dilemas. Moral e interés coincidirían. Y esto, de nuevo, deja sin papel a la moral.

Pero tanto los dilemas como el enfrentamiento entre moral e interés propio existen. En lo que resta nos dedicaremos a examinar este enfrentamiento y el papel de la moral en la solución de los dilemas.

CAPITULO IX

LA MORAL COMO PUNTO DE VISTA

El último texto citado en el capítulo anterior puede ser interpretado al margen de cualquier relación con el intento de derivar la moral del propio interés. Así entendido, el texto nos diría que un individuo justo es aquel que tiene una capacidad afectiva para la moralidad, lo cual significa que ese individuo, una vez que reconoce un curso de acción como justo, se encontrará a sí mismo, por esta única razón, dispuesto a adoptarlo. Es decir, estará dispuesto a seguirlo independientemente de las consecuencias que esto tenga para él.

Quando todos los agentes que intervienen en una determinada situación, o al menos un número suficiente de ellos, tiene esta capacidad y esta disposición, los dilemas se resuelven del modo indicado en las soluciones 4 y 5 de nuestro cuadro¹. Sin embargo, hay que señalar dos cosas

1. Como se mostró en el capítulo anterior, esto es incompatible con el intento de basar la moral en el interés propio. Más bien al contrario, un individuo con tal capacidad debe estar dispuesto a juzgar sobre lo justo con independencia de lo que eso signifique para él. Esto muestra que la moral es un punto de vista distinto del punto de vista del interés propio, y que, por consiguiente, puede en ocasiones dictar una conducta distinta.

2. Sin embargo, tal y como lo muestra el hecho de que la generalización de esa disposición posibilita la solución de los dilemas, debe haber algún tipo de conexión entre moral y racionalidad individual.

La tarea que nos ocupará en adelante será por tanto doble. En este capítulo intentaremos caracterizar el punto de vista propio de la moral y más adelante trataremos de establecer la conexión entre moral e interés propio y de aclarar la naturaleza de tal conexión.

¹ Ver p. 136

Quando en el capítulo 8 hablábamos del altruismo como una característica cuya posesión por parte de los agentes puede hacer que estos, en una situación de dilema, elijan la estrategia A, mencionamos por primera vez la distinción entre intereses personales e intereses de otro tipo, a saber, intereses que el agente tiene sólo si se obliga a juzgar la situación y realizar su elección desde un punto de vista moral. Esta distinción corresponde a la establecida por Harsanyi entre preferencias personales y preferencias morales. Para ver que es lo que distingue unas de otras empezaremos por analizarlas separadamente.

Las preferencias personales de un individuo son "sus preferencias reales, típicamente basadas en sus propios intereses personales y en los intereses de aquellos que le son más cercanos"². Las preferencias personales o subjetivas de un individuo son sus preferencias en el "sentido pleno de la palabra", e.d., las preferencias que cada individuo tiene realmente y que determinan su función de utilidad. Son, por tanto, las preferencias de las que hemos hablado hasta ahora.

Un error habitual, contra el cual ya hemos argumentado anteriormente, consiste en entender que las preferencias personales son por definición preferencias egoístas. Esto no es así. Se puede tener una preferencia personal basada únicamente en intereses egoístas, pero también puede tenerse una preferencia personal basada en intereses altruistas. Hay casos en los que es claro que un determinado interés es egoísta, como sucede cuando se tiene un interés en uno mismo, al igual que hay casos de intereses indudablemente altruistas, como cuando alguien se interesa por lo que le sucede a un extraño. Pero también hay casos fronterizos que dudaríamos en calificar de una u otra manera, como sucede con el interés que una persona demuestra por sus padres o sus hijos. Sin embargo, con el fin de distinguir claramente un interés personal altruista de un interés moral, conviene definir de un modo algo más preciso un interés egoísta. Reservaremos el nombre de "interés egoísta" a aquel que una persona tiene por sí mismo, y llamaremos "interés altruista" al interés que una persona siente por algún otro. Hablando de este modo, diremos que el interés de una madre por su hijo es altruista. En este sentido, "egoísta" es sinónimo de "interesado en uno mismo" y "altruista" de "no interesado en uno

² Harsanyi (76), p.IX.

mismo", lo cual recoge en buena medida el uso habitual de estos términos. Resulta claro ahora que las preferencias personales no están necesariamente, ni siquiera típicamente, basadas en intereses egoístas.

Las preferencias personales son las utilizadas en la definición del agente racional y en la teoría de la racionalidad práctica que hemos analizado. Es, por tanto, el comportamiento racional de los agentes basado en estas preferencias el que da lugar a la aparición de los dilemas. Sin embargo, puede pensarse que hay un tipo de intereses personales, a saber, los intereses altruistas, que dan lugar a un solución de los dilemas. Es necesario hacer aquí una matización y una distinción.

La matización consiste en precisar qué tipo de altruismo resuelve los dilemas. Los participantes no deben ser altruistas puros, e.d., personas que no tienen *ningún* interés por sí mismas, ya que de este modo se plantearía un dilema parecido.

Este dilema surgiría entre dos altruistas puros cuando se encuentran en situación de elegir entre a) procurarse a sí mismo un gran beneficio o b) proporcionar al otro un beneficio más pequeño. Un altruista puro elegiría b, con lo cual el resultado sería peor para todos. Por ejemplo, consideremos la siguiente matriz

	B ₁	B ₂
A ₁	(1,1)	(0,9)
A ₂	(9,0)	(8,8)

Ejemplo 10

Cada uno de los jugadores puede a) darle 1 al otro o b) darse 8 a sí mismo. Llamaremos A a la estrategia que da 1 al otro (e.d., las estrategias A₁ y B₁) y S a la que da 8 a uno mismo (A₂ y B₂). Si ambos eligen A, el resultado será (1,1). Pero hay otro resultado que los dos preferirían a este, a saber (8,8). Supongamos que los dos se ponen de acuerdo en no ayudarse y elegir ambos la estrategia B, con el fin de obtener ese resultado. Pero, de nuevo, este acuerdo no es estable. El jugador uno razonaría así. Si el jugador 2 mantiene el acuerdo y hace B, entonces es mejor para él que yo haga A. Y si no lo mantiene y hace A, entonces también es mejor para él que yo haga A. De modo que haré A. El jugador 2 haría un razonamiento similar y llegaría a la misma conclusión. Por tanto, sería

(1,1) el resultado.

Parece por tanto que la transformación de los agentes en altruistas puros no solucionaría los dilemas. Más bien, la solución deberá consistir en que los agentes sean lo suficientemente altruistas, no sobrepasando en ningún caso el límite de la benevolencia imparcial, e.d., una consideración imparcial de los intereses de todos, incluidos los propios. Por "altruismo" deberá entenderse "suficiente altruismo", ya que unos altruistas puros se enfrentarían a dilemas paralelos a los de unos agentes no altruistas.

Pero además es necesario precisar el alcance de este altruismo. Pensemos de nuevo en el caso de los prisioneros. Como ya sabemos, si los prisioneros son egoístas surgirá el dilema. Imaginemos ahora que sucedería si el juego se desarrollara no entre los propios prisioneros sino entre sus respectivas esposas. Según nuestra definición, sus intereses no serían egoístas, sino altruistas. Sin embargo, es fácil comprobar que entre ellas surgiría el mismo dilema. Esto sucede porque sus intereses altruistas están dirigidos a personas distintas, lo cual hace que sus intereses se enfrenten como sucede en el dilema originario. Llamaremos a este altruismo dirigido a personas concretas *altruismo particular*. Este tipo de altruismo puede solucionar algunos dilemas concretos, pero no es una solución de carácter general. Por ello, este tipo de solución sería lo que en su momento calificamos de "solución psicológica", que se distinguía de las llamadas "soluciones morales" precisamente por su carácter particular. En contraposición a este tipo de altruismo está el *altruismo universal*.

Pero tampoco el altruismo universal garantiza la solución de los dilemas. Veamos un ejemplo. En un colegio deciden contratar un profesor para ofrecer a los alumnos alguna materia extraescolar. El presupuesto de la escuela sólo alcanza para contratar un profesor cinco horas a la semana. Algunos profesores son partidarios de contratar a un profesor de ballet y otros de contratar un profesor de kárate. Los intereses de ambos grupos son altruistas. Todos están interesados en que los niños tengan la mejor formación posible. El único problema es que no están de acuerdo en qué es lo mejor posible. Después de mucho discutir, llegan a un acuerdo. Contratarán dos profesores y cada uno dará, en lugar de cinco horas de clase semanales, dos. Uno de los partidarios del kárate, a quien llamare-

mos K, queda encargado de hablar con su prima, bailarina y con buenos conocimientos de la materia, mientras que uno de los amantes del ballet, al cual nos referiremos con el nombre de B, se encarga de localizar a un profesor de kárate, ya que al lado de su casa hay una academia de artes marciales especialmente buena. En esta situación, aparece el dilema. ¿Cumplirán K y B el acuerdo? Pongámonos en la situación de B. Si B no cumple el acuerdo y no busca un profesor de kárate, y K por su parte sí lo cumple, el resultado es maravilloso para los niños. Tendrán un profesor de ballet que les dará clase todos los días. Y si K no cumple con su parte, entonces es mejor para los niños que tampoco él lo haga, porque, de hacerlo, las pobres criaturas tendrían clase diaria de kárate, lo cual sería sin duda horrible para ellos ya que se convertirían en una especie de Rambo orientalizado. Por su parte, K haría un razonamiento paralelo, a resultados de lo cual los niños acabarían por no tener ninguna clase. Pero esto es peor que el resultado del acuerdo, según el cual los niños tendrían al menos alguna clase de lo que deben aprender (naturalmente, por su bien). Llamaremos a este tipo de altruismo altruismo paternalista. En contraposición a este altruismo está el altruismo a secas, según el cual una persona tiene en cuenta los intereses de los demás, siendo los propios interesados los que determinan cuáles son sus intereses. Hasta aquí las matizaciones.

Parece que las preferencias personales basadas en el altruismo universal proporcionan un medio de resolución de los dilemas. Veamos si es así. Pensemos de nuevo en el dilema de los maestros, pero supongamos que ahora las preferencias personales de todos ellos son lo que hemos llamado altruistas a secas. De modo que para decidir que hacer con su presupuesto para actividades extraescolares convocan una asamblea, presentan a los niños las distintas alternativas disponibles y les piden que voten a favor de una de ellas. Supongamos que del resultado de la votación se desprende que parte de los niños quiere tomar clases de kárate, mientras que otra parte prefiere tomar clases de ballet, y supongamos, para simplificar, que esta diversidad de opiniones agrupa a los niños en dos grupos de igual número. Supongamos además que todos los niños prefieren tener alguna clase de lo que les gusta a no tener ninguna. Para los profesores el orden de preferencia entre las distintas alternativas está determinado únicamente por las preferencias de los niños. Debido a la

división de opiniones de los niños y a lo limitado del presupuesto, la posibilidad de que todos los niños obtengan la alternativa preferida (cinco horas de kárate unos, y cinco horas de ballet otros) sencillamente no existe. Por ello, los profesores deciden contratar a un especialista en cada una de las materias por dos horas semanales cada uno, con lo cual todos los niños verán satisfecha al menos la segunda alternativa de su orden de preferencias (alguna clase de su materia favorita). Uno de los profesores se encarga de contratar a un profesor de kárate y otro a uno de ballet. La situación de elección en este caso es similar a la anterior, pero existe entre ambas una diferencia sumamente importante. Mientras que en el ejemplo originario teníamos un caso de juego de intereses mixtos, ahora la situación queda adecuadamente representada como un juego de intereses idénticos. Por tanto, la pregunta de si los profesores cumplirán su parte es retórica. Y esto sucede porque, mientras que en el caso anterior y, en general, en todos los casos que presentan un dilema, nos encontramos con una situación en la que intervienen varios jugadores *cada uno con una función de utilidad y un orden de preferencias*, que sólo coinciden parcialmente, en el caso de altruistas universales no paternalistas la función de utilidad y el orden de preferencias de todos ellos *es el mismo*.

¿Significa esto que cuando las preferencias personales están determinadas por intereses altruistas del tipo adecuado los dilemas se solucionan? En realidad no. Una cosa es solucionar un problema y otra verse libre de él. Para solucionar un problema primero es necesario que este exista. Pero uno puede verse libre de un problema sencillamente porque este no se plantea, porque nunca llega a existir. Y este es lo que sucede en este caso, a saber, que el dilema no se resuelve, pero no porque quede sin solución, sino porque no hay nada que resolver.

Parfit afirma que las soluciones 3 y 4 representan la abolición de los dilemas mientras que la solución 5 resuelve el problema práctico dejando el problema teórico intacto³. Esto no me parece exacto. El caso que acabamos de discutir, e.d., el caso en el que las preferencias perso-

³ Ver cuadro en p.136

nales de los agentes están determinados por intereses altruistas⁴, da lugar a la llamada solución 3. Pero hablar aquí de "solución" es algo que sólo puede hacerse en sentido impropio. Esta característica de las preferencias personales de los agentes, al igual que otras como tener en alta estima el propio honor etc, no soluciona ningún dilema. Más bien, imposibilita su aparición. La solución 3 no es una solución de nada⁵. Todo lo que ocurre es que los agentes eligen A porque, debido a las características de sus preferencias personales, A no es peor para ellos.

Sin embargo, en multitud de ocasiones surgen los dilemas. El altruismo universal es raro y el altruismo no paternalista, rarísimo. No habría dilemas si fuéramos de otro modo, pero somos como somos. Esto no quiere decir que no podamos cambiar. De hecho, cambiamos. Pero, desgraciadamente, modificar nuestras preferencias personales en la dirección del altruismo es sumamente difícil, prácticamente imposible. Afortunadamente, no es necesario. Los dilemas pueden solucionarse. Es posible que los dilemas surjan debido a las características de nuestras preferencias personales y que, a pesar de ello, puedan solucionarse mediante la elección basada en otro tipo de preferencias, a saber, las preferencias morales.

Las preferencias morales de un individuo son "sus preferencias hipotéticas, que tendría si se forzara a sí mismo a juzgar el mundo desde la moral"⁶. Por su carácter hipotético, las preferencias morales sólo son preferencias en un sentido cualificado del término. No son, como las preferencias personales, las que un individuo tiene de hecho, sino las que podría tener si adoptara un determinado punto de vista. Las preferencias morales sólo expresan lo que un individuo prefiere en los momentos en los que este juzga desde ese punto de vista.

Que las preferencias morales se definan como hipotéticas no signi-

⁴ En adelante, por "altruismo" entenderemos altruismo universal no paternalista.

⁵ Hay un sentido en el que la solución tres es una auténtica solución, y es sin duda a este al que Parfit hace referencia. Cuando existe un dilema, uno de los modos en los que este se puede hacer desaparecer consiste en cambiar las preferencias personales de los agentes en alguno de los sentidos indicados, como por ejemplo, haciendo que sus intereses pasen a ser altruistas. Pero una vez que contamos con agentes altruistas, los dilemas dejan de plantearse.

⁶ Harsanyi (76), p. IX.

fica que no puedan tenerse realmente o que, de tenerlas, pasen a ser necesariamente preferencias personales⁷. Lo único que significa es que, en determinadas circunstancias, nuestras preferencias personales serían preferencias morales, en el sentido de preferencias surgidas de una consideración imparcial de los intereses de todos⁸. El motivo por el que se definen como hipotéticas es que "preferencia personal" es un término básico de la teoría, a partir del cual se define la función de utilidad de un individuo, mientras que "preferencia moral" no lo es y debe por tanto, ser introducido en la teoría poniendolo en relación con los términos propios de esta.

Introducir el término "punto de vista" al hablar de preferencias morales puede hacer surgir una confusión. En efecto, puede pensarse que las preferencias morales están ligadas a un punto de vista determinado, mientras que las preferencias personales surgen sin necesidad de adoptar un punto de vista. Esta confusión se apoya en la creencia de que los juicios pueden realizarse o bien desde un determinado punto de vista o bien desde ninguno, o lo que es lo mismo, desde un punto de vista "objetivo" o "natural" que, por ello mismo, ya no es un punto de vista. Algo así como si una habitación pudiera ser contemplada desde varios ángulos y también desde ninguno. Pero al igual que esto es imposible también lo es el realizar un juicio sin hacerlo desde algún punto de vista. Una preferencia moral es la que se tiene si se adopta el punto de vista de la moral. Pero también una preferencia personal es la que se tiene al adoptar un determinado punto de vista, a saber, el punto de vista del interés personal. Ambos tipos de preferencia expresan un punto de vista.

Sin embargo, hay un sentido en el que las preferencias personales son independientes de un punto de vista, y es en este sentido en el que es posible definir las sin hacer mención a la adopción de un punto de vista, cosa imposible en el caso de las preferencias morales. Lo que sucede es que las preferencias personales expresan un punto de vista "natural"

⁷ De hecho, Harsanyi propone una definición alternativa de las preferencias morales en la que esto resulta claro: "también podemos decir que representa sus preferencias solo en aquellos momentos - posiblemente poco habituales- en los que se fuerza a una especial actitud moral imparcial sobre sí mismo." (Harsanyi, 77c, p.50)

⁸ Ver Nota 8 más adelante.

en el sentido de ser el punto de vista propio de las preferencias. Esto es lo que hace que las preferencias personales sean preferencias en el sentido pleno del término y lo que las confiere su carácter real a diferencia del carácter hipotético de las preferencias morales. Una ordenación de un conjunto bajo la relación de preferencia es una ordenación según lo que el agente prefiere de hecho. Expresa precisamente lo que prefiere un individuo. En contraste, la preferencia moral ordena las alternativas no según las preferencias que tiene un agente sino *las que tendría* si adoptara un punto de vista que habitualmente no adopta para establecer una relación de preferencia. El punto de vista propio de las preferencias es el punto de vista personal.

Es importante tener presente que tanto las preferencias personales como las morales son preferencias. Es decir, ambos tipos de relación establecen entre un conjunto de alternativas una relación de ordenación a partir de la cual se hace posible una elección racional. Podemos preguntarnos ahora si la relación de preferencia personal y la de preferencia moral tienen el mismo campo de aplicación, e.d., si los elementos de un conjunto ordenado según la relación de preferencia personal pueden también ser ordenados según la relación de preferencia moral y viceversa.

Podemos contestar esta pregunta inquiriendo en primer lugar cual es el campo de aplicación de la relación de preferencia personal. En el capítulo II caracterizamos una ordenación diciendo, entre otras cosas, que era completa, e.d., que era una relación entre todos los miembros del conjunto. Por tanto, puesto que la relación de preferencia da lugar a una ordenación debemos entender que esta relación puede establecerse entre dos elementos cualesquiera del conjunto. Sin embargo, esto no es decir mucho. Pensemos en la relación "ser al menos tan alto como" aplicada al conjunto de edificios del mundo. Es evidente que esta relación es completa en este caso, pues de cualesquiera dos miembros de este conjunto puede decirse cual es más alto o si son igual de altos. Supongamos ahora que el conjunto que queremos ordenar es el de la herencia cultural española del siglo XVII. Este conjunto no puede ordenarse según la relación "ser al menos más alto que", puesto que si bien esta relación puede establecerse entre algunos miembros del conjunto (por ejemplo entre los edificios, esculturas y pinturas) no puede establecerse en todos los casos (el Panteón de los Reyes de El Escorial no es ni más alto ni menos ni igual que

El Quijote).

Este ejemplo muestra la necesidad de precisar sobre qué conjunto es completa una relación. La relación que nos interesa es la de "ser preferido o indiferente a". Esta relación se parece a la de "ser al menos tan alto como" en que ambas establecen un orden jerárquico que admite empates. Esto significa que ambas agotan todas las posibilidades de ordenación de un conjunto respecto a la propiedad a la que hacen referencia, e.d., respecto a la altura en un caso y a la deseabilidad en otro. Por ello ambas relaciones son completas cuando se aplican a un conjunto de elementos que poseen esa propiedad. Respecto a cualquier conjunto de elementos que tengan altura, podrá decirse de cualquier par x e y de elementos del conjunto si x es al menos tan alto como y o si y es al menos tan alto como x . Esta relación es completa respecto al conjunto de todas las cosas físicas. De modo similar, la relación "ser preferido o indiferente a" es completa cuando se aplica sobre el conjunto de cosas susceptibles de ordenarse según la preferencia. Este conjunto es el de todas las alternativas posibles.

Ahora bien, lo que nos interesa es saber si la relación de preferencia moral es completa respecto al mismo conjunto para el que es completa la relación de preferencia personal. Esto equivale a preguntar si hay algún conjunto de elementos que no puedan ordenarse según la relación de preferencia moral. La respuesta a esta pregunta es negativa. La única objeción posible a esta respuesta consiste en argumentar que hay situaciones sin relevancia moral alguna. A pesar de que esto es cierto, no constituye una buena objeción, pues estos casos son en realidad casos de indiferencia moral.

Tenemos por tanto dos relaciones distintas que aplicadas a los mismos conjuntos dan lugar a ordenaciones. Esto es, un conjunto cualquiera de alternativas puede ser ordenado mediante la aplicación de dos relaciones distintas. Es por ello posible que un elemento del conjunto ocupe una posición superior en el orden jerárquico establecido según una de estas relaciones y una inferior según la otra. Sea R la relación de preferencia personal y M la relación de preferencia moral. Es posible que tomados dos elementos x e y suceda xRy y también yMx . Esto es lo que quiere decirse al afirmar que la moral y el interés propio pueden entrar en conflicto. La posibilidad de este conflicto es lo que hace posibles

las soluciones 4 y 5 a los dilemas, pues estas se dan precisamente cuando los agentes eligen A con indiferencia de que A sea o no mejor para ellos. Es posible elegir A sin que esto sea mejor para el agente si la elección se realiza a partir de una ordenación establecida según la relación de preferencia moral. En adelante diremos que estas soluciones se dan cuando los agentes eligen A porque A es la estrategia que consigue el mejor resultado desde el punto de vista moral.

Las preferencias morales son aquellas preferencias que el agente tendría si se forzara a juzgar el mundo desde un punto de vista moral. El siguiente paso será por tanto caracterizar este punto de vista. Hasta ahora lo único que sabemos es que el punto de vista moral es distinto del punto de vista del interés personal. Esto no quiere decir que la moral y el interés propio siempre estén en conflicto. Hay ocasiones en las que lo mejor desde el punto de vista moral es también lo mejor desde el punto de vista del interés personal. Más bien lo que quiere decir es que son independientes y que, por ello, pueden no coincidir. El punto de vista del interés personal es esencialmente parcial. En contraste, el punto de vista moral se caracteriza por la imparcialidad. Es esencialmente un punto de vista imparcial e impersonal.

A pesar de ser esencial, la caracterización del punto de vista moral como imparcial no resulta suficiente. Es necesario precisar acerca de qué se es imparcial. En este sentido, y en tanto que la moral es una ordenación sobre las preferencias alternativa a la del interés propio, podemos decir que la imparcialidad se ejerce respecto a los distintos intereses de los distintos agentes. Es decir, una elección realizada desde el punto de vista moral sería siempre igual a la realizada por un altruísta universal.

La cuestión fundamental es cómo es posible que un agente racional en el sentido habitual realice una elección desde el punto de vista moral. No basta con decir a qué sería equivalente esa manera de elegir. Más bien lo que se necesita es una explicación de cómo y en qué circunstancias sería posible que un agente racional realizara una elección impar-

cial". La respuesta más directa consiste en determinar qué es lo que hace que las elecciones de un agente racional sean parciales. Sin duda, la parcialidad deriva del hecho de que cada agente racional tiene una función de utilidad e intenta mediante sus elecciones maximizar su propia utilidad individual. Sabe qué es lo que tiene y qué es lo que puede esperar de una acción determinada. Harsanyi utiliza un ejemplo especialmente claro

"Supón que alguien nos dice: "yo prefiero con mucho nuestro sistema capitalista a cualquier sistema socialista porque resulta que bajo el sistema capitalista yo soy un millonario y llevo una vida muy satisfactoria, mientras que, con toda probabilidad, bajo un sistema socialista yo sería como mucho un funcionario de poca monta y mal pagado." Esto puede ser un juicio de preferencias personales muy razonable desde su propio punto de vista individual. Pero nadie lo llamaría juicio de valor *moral*, porque sería obviamente un juicio basado primordialmente en el interés propio".¹⁰

En efecto, lo que hace posible un juicio parcial como este es el hecho de que el agente sabe perfectamente lo que representa para él un sistema capitalista. Sabe la situación que ocupa en una sociedad ordenada de ese modo y conoce las probabilidades que tendría en una sociedad socialista.

Las elecciones que realizamos habitualmente, basadas en preferencias personales, se apoyan en el conocimiento de estos datos. Es más, cuanto más preciso sea nuestro conocimiento en este sentido más fácil será que hagamos una buena elección. Naturalmente, una buena elección parcial. Por tanto, puede suponerse que un agente racional realizaría una elección imparcial si ignorara estos datos. Supongamos que yo tengo que partir un pastel en cuatro partes, una de ellas para mí y las otras para otras tres personas, a las que las gusta el pastel tanto como a mí y por las cuales yo no siento el menor interés. La cuestión no es sólo en qué circunstancias yo partiría el pastel de modo imparcial, sino en qué circunstancias esta partición resultaría racional. De las cuatro personas una elegirá un trozo en primer lugar, otra en segundo etc. El orden de

* Puede que alguien dude de la necesidad de partir del supuesto de unos agentes racionales mutuamente desinteresados. Sin embargo, como ya se ha discutido anteriormente, el hecho de que los hombres tengan distintos intereses e intenten satisfacerlos en la mayor medida posible es una de las circunstancias que hacen posible y necesaria la moral. Por ello es necesario tratar de caracterizar la moral tomando como sujeto al agente racional tal y como hasta ahora ha quedado caracterizado.

¹⁰ Harsanyi (77), p.631

elección puede establecerse según diversos criterios. Supongamos que el criterio es contar el chiste más divertido. Entonces, si yo ignorara el ingenio de los otros tres así como el mío propio, yo no podría hacer ningún cálculo acerca del posible puesto en el que me va a tocar escoger. Sin duda, en esta situación lo más racional sería que partiera el pastel a partes iguales.

Generalmente se admite que una elección sería imparcial si el agente efectuara su elección en una situación de incertidumbre respecto a cómo le afectará personalmente el resultado de su acción. La elección imparcial se caracteriza como la que un agente realizaría si estuviera en una posición especial de ignorancia. En esta situación especial de incertidumbre lo que ignoraría un agente sería su propia identidad. Los agentes ignorarían básicamente sus capacidades físicas y mentales, su posición social y económica y sus preferencias personales concretas. Todo el conocimiento con el que contarían para realizar su elección sería de carácter general. Respecto a ellos mismos sabrían que tienen unas preferencias determinadas (aunque no saben qué preferencias exactamente) y que los demás también las tienen, y tendrían un conocimiento general completo acerca de la naturaleza de la sociedad. En resumen, tendrían todo el conocimiento de carácter general y ninguno acerca de lo que les hace individuos. Sabrían que son individuos, pero no qué individuo son. La elección realizada en unas condiciones de este tipo sería necesariamente imparcial y, en tanto que imparcial, sería una elección realizada desde el punto de vista moral¹².

¹² A pesar de que esta caracterización del punto de vista moral es generalmente aceptada, distintos autores difieren en lo tocante a cual sería la conducta racional de un agente en estas condiciones de ignorancia y presentan distintos modelos para caracterizar en qué consiste elegir desde un punto de vista moral. Por ejemplo, los modelos presentados por Rawls y Harsanyi coinciden en caracterizar la elección moral como aquella que se realizaría tras el "velo de la ignorancia", pero difieren en la regla de decisión adecuada para tales situaciones. En concreto, mientras que Rawls sostiene que la regla de decisión adecuada es la regla maximin, Harsanyi afirma que un agente racional que se encontrara tras el velo de la ignorancia elegiría aquella alternativa que maximizara la utilidad media aritmética de los individuos afectados por la elección.

CUARTA PARTE

POR QUE ADOPTAR EL PUNTO DE VISTA
MORAL

Esta pregunta, de la cual nos ocuparemos en esta parte de la tesis, debe ser contestada en términos de interés propio. El motivo es, simplemente, que lo que queremos saber es por qué un agente racional, tal y como ha sido caracterizado en capítulos anteriores, ha de adoptar el punto de vista moral.

Para encontrar una respuesta a nuestra pregunta, partiremos de dos hechos indiscutibles. En *primer* lugar, salvo que se den determinadas circunstancias, tales como reciprocidad, capacidad de reconocimiento mutuo y posibilidad de aplicar sanciones, actuar desde el punto de vista moral o, dicho en la terminología que utilizaremos con más frecuencia en adelante, elegir la estrategia A (truista), es individualmente irracional. Esto es lo que aprendemos de las situaciones de dilema. Cada uno de los agentes hará lo mejor si escoge la estrategia B. Es decir, hagan lo que hagan los demás, cada uno obtendrá un resultado mejor si él hace B. Pero, en *segundo* lugar, hay otra cosa que aprendemos de los casos de dilemas, a saber, que *son dilemas*. Es decir, a pesar de que cada uno hace lo mejor que puede hacer haciendo B, sucede que, si todos hacen lo que es mejor para cada uno, el resultado es peor para todos que si cada uno hiciera lo que es peor para él. Nos referiremos a este hecho diciendo que, aunque la elección de B en esas situaciones es *individualmente racional*, es *colectivamente irracional*. De modo inverso, la elección de la estrategia A resulta *ser colectivamente racional*, a pesar de *ser individualmente irracional*.

Teniendo en cuenta estos hechos, hay dos posibles respuestas a nuestra pregunta

1. La primera posibilidad es negar, o al menos no tener en cuenta, la distinción entre racionalidad individual y racionalidad colectiva, y afirmar que la única racionalidad es la racionalidad individual. Desde esta postura, el único motivo que un agente racional puede tener para adoptar el punto de vista moral es el deseo de obrar moralmente.

2. Otra posibilidad es admitir que, junto con la racionalidad individual, puede hablarse de racionalidad colectiva. Esto hace posible la pregunta acerca de la relación ^{entre} ambos tipos de racionalidad. Para el análisis de estas relaciones puede utilizarse otra distinción, a saber, la distinción entre racionalidad directa y racionalidad indirecta, y la consideración de la moral, e.d., de la conducta colectivamente racional, como una conducta indirectamente racional en el nivel individual.

En el capítulo X nos ocuparemos de la caracterización de la racionalidad directa y la racionalidad indirecta. En el capítulo XI analizaremos las características de nuestra teoría que hacen que sea auto-frustrante a nivel colectivo y consideraremos la posibilidad de una teoría alternativa sobre la racionalidad práctica que no lo sea. Por último, tras haber analizado estas cuestiones preliminares, intentaremos dar una respuesta satisfactoria a nuestra pregunta.

CAPITULO X. RACIONALIDAD INDIRECTA

I. *Racionalidad directa y Racionalidad indirecta.*

Supongamos que me encuentro en un apuro. Es de noche, estoy lejos de mi casa y he perdido la cartera. En ese momento veo pasar un taxi. Si le prometo al taxista que le mandaré el dinero por la mañana, me llevará a casa. Supongamos además que somos transparentes, es decir, que los demás pueden leer nuestras disposiciones de un modo inequívoco. De este modo, el taxista sabe si yo soy el tipo de persona que cumple sus promesas o no lo soy. Supongamos que yo soy un individuo racional. Lo mejor que puedo hacer es prometerle al taxista que le pagaré. Pero, una vez que he llegado a mi casa, lo mejor para mí será incumplir mi promesa. Como el taxista sabe esto, no creerá mi promesa, con lo que el resultado será para mí el peor posible. ¿Qué puedo hacer en este caso? Desde luego, no puedo realizar la promesa con intención de romperla. Pero puedo intentar hacer la promesa totalmente en serio, con toda la intención de cumplirla. Sin embargo, esto no me serviría tampoco de mucho. El taxista sabe que soy una persona que siempre actúa maximizando la utilidad, de modo que, dado el tipo de persona que soy, no voy a pagarle. Parece que sólo puedo hacer dos cosas. Puedo hacer algo para alterar la situación en la que me encontraré al llegar a casa. Por ejemplo, puedo dejarle en prenda algo lo suficientemente valioso como para que lo mejor para mí sea volver a pagarle. O, alternativamente puedo convertirme a mí misma en una persona que cumple sus promesas sin tener en cuenta si esto es mejor o peor para ella. Llamemos R a la primera solución e I a la segunda.

Consideremos otro caso. Un ladrón entra en mi casa. Mientras fuerza la puerta, me oye llamar a la policía, de modo que el ladrón, teniendo en cuenta la distancia que separa mi casa del puesto de policía más cercano, puede calcular el tiempo que tiene antes de que lleguen. En vista de esto, me amenaza con que, a menos que le dé todo el dinero en cinco minutos, comenzará a matar a mis hijos uno por uno. La situación es realmente desesperada. Si no le doy el dinero mis hijos morirán. Si se lo doy, puesto que tanto mis hijos como yo le hemos visto la cara, probablemente nos matará también a todos para que no podamos reconocerle. Estoy perdido. Soy un agente racional que trata de conseguir el mejor resultado posible. Pero en esta situación, parece que haga lo que haga el resultado

será horrible.

Supongamos que tengo a la mano una droga que tiene el poder de volver irracional a quien la tome por un corto periodo de tiempo. Si la tomo, a los pocos segundos será evidente que me he vuelto loco. Por ejemplo, empezaré a preferir lo peor. Le diré al ladrón que, puesto que quiero a mis hijos, me parece estupendo que los mate. Si me tortura, le diré que, dado que el dolor es insoportable, estoy encantado de que continúe. El ladrón queda automáticamente desarmado. Como además, dado el estado en que me encuentro, es muy improbable que pueda reconocer a nadie, el ladrón ya no teme que pueda reconocerle. El único inconveniente es que en mi irracionalidad puedo causarles algún daño a mis hijos o a mi mismo. Pero como la policía tardara poco en llegar, el riesgo es mínimo o, al menos, menor que el que corro si continuo siendo racional¹. Esta sería una posible solución a mi problema, a la que llamaremos I. Pero puede que haya aún otra solución, bastante parecida a la anterior pero distinta en aspectos sustanciales. Supongamos que tengo otra droga, a la que llamaremos droga(2), que tiene el poder de alterar la ordenación de las preferencias de quien la tome. Si la tomo, el resultado será que perderé todo el interés que tengo por mis hijos y por mí mismo y adquiriré un inmenso amor por el atracador. Una vez que mis preferencias se hayan alterado de ese modo, estaré sumamente dispuesto a darle al ladrón todo lo que me pida, y no le denunciaré mientras dure el efecto de la droga. El ladrón verá que causarme daño a mí o a mis hijos ni le serviría de nada ni sería tampoco necesario. Además, tampoco tiene que temer que pueda identificarle más tarde. Mientras dura el efecto de la droga el tiene tiempo más que suficiente para ponerse a salvo. Por supuesto, existe el riesgo de que le puedan coger a pesar de todo. Pero este riesgo puede merecer la pena, si es suficientemente bajo. La posibilidad de que le cojan existe de todos modos. No hay crímenes perfectos. Si mi identificación tardía no aumenta demasiado esta posibilidad, el aumento de riesgo queda compensado con la menor gravedad del delito. Llamaremos a esta solución "solución R".

¿Qué es lo que muestran estos casos? Fijémonos en primer lugar en las soluciones R de ambos ejemplos. Al tratar estos casos, Parfit no tie-

¹ Ambos ejemplos están extraídos de Parfit (86), pp.7 y 13 respectivamente.

ne en cuenta estas posibles soluciones, limitándose a tratar las soluciones I. El motivo es evidente. Las soluciones R no plantean ningún problema teórico a nuestra teoría de la racionalidad. Más bien, son soluciones que podríamos llamar "internas" a la teoría. En ellas, el agente no tiene necesidad de dejar de ser un individuo racional. Lo único que tiene que hacer es alterar la situación de tal modo que la acción maximizadora pase a ser otra. En el primer ejemplo, antes de producirse el cambio, si soy racional no pagaré al taxista. Pero precisamente por esto el resultado será para mí el peor. Al dejarle una prenda valiosa, lo que hago es alterar la situación de tal modo que cuando llega la hora de realizar la elección, esto es, cuando llego a mi casa, la acción maximizadora ya no es no pagar, sino pagar y recuperar la prenda. Esta alteración de la situación tiene como resultado algo mejor para mí, a saber, que el taxista me llevará a casa. Algo similar ocurre en el segundo ejemplo. Antes de tomar la droga(2) e invertir las preferencias, la acción maximizadora conduce a un resultado altamente indeseable². Sin embargo, una vez que he modificado mis preferencias, la acción maximizadora consiste en acceder a las peticiones del ladrón y en no denunciarle. Y esta acción resulta también maximizadora con respecto a mis preferencias originales, siendo así el mejor resultado también cuando yo recupere mi estado normal. Habré perdido mi fortuna, pero mis hijos y yo seguiremos vivos. Si mis hijos son muy pequeños pueden padecer secuelas psicológicas producidas por la impresión de que yo no les considere importantes en esa ocasión, pero al menos estarán vivos y las terapias hacen milagros.

Las soluciones R pertenecen a una categoría de lo que podríamos llamar "trucos" tremendamente habituales. Este tipo de estrategias son similares a las recogidas por Elster bajo el nombre de *Racionalidad im-*

² En realidad, tal y como presenta el caso Parfit, puede parecer que lo que ocurre es que no hay acción maximizadora en absoluto, en el sentido de que, haga lo que haga, el resultado será exactamente el mismo. Sin embargo esto no es así necesariamente. Si le doy el dinero, nos estará con toda probabilidad, pero si no se lo doy, existe la posibilidad de que pierda los nervios, o de que la policía llegue algo antes de lo esperado, a tiempo de salvar alguna vida, o de que alguien pase por allí casualmente. La ventaja de este segundo curso de acción consiste precisamente en que da más tiempo a que ocurra lo improbable. Si esto no es muy convincente, aún puede pensarse que, si no accedo a sus pretensiones, aunque nos mate a todos igualmente, al menos no se llevará el dinero. Como puede verse, aunque la diferencia de utilidad entre un resultado y otro no sea muy grande, es lo suficiente como para poder hablar de "acción maximizadora".

*perfecta*³. En este tipo de casos el agente consigue el mejor resultado por métodos indirectos. La solución R del primer caso ilustra lo que para Elster es el método privilegiado para lograr soluciones indirectas. El ejemplo paradigmático de este método es la estratagema empleada por Ulises, quien, sabiendo que no podría resistir la tentación de las sirenas, pidió a sus compañeros que le ataran, de modo que no pudiera caer en la tentación. En recuerdo de este caso paradigmático, Elster bautiza este método con el nombre de "atarse uno mismo". Para que una estratagema pertenezca propiamente a este método, debe cumplir las siguientes características:

1. Atarse uno mismo es llevar a cabo cierta decisión en un momento t_1 con el fin de incrementar la probabilidad de llevar a cabo otra decisión en un momento posterior t_2 . El cambio esperado de la probabilidad de la acción posterior debe ser el motivo para la primera acción.
2. Si la primera de las dos acciones tiene como efecto inducir un cambio en el conjunto de opciones que serán viables en t_2 , entonces la acción realizada en t_1 no será de atarse uno mismo si el nuevo conjunto incluye el conjunto primitivo.
3. La decisión llevada a cabo en t_1 debe tener el efecto de establecer un mecanismo causal en el mundo externo.
4. La resistencia a llevar a cabo la decisión en t_1 debe ser menor a la resistencia que hubiera existido para llevar a cabo la decisión en t_2 si no se hubiera realizado la acción en t_1 .
5. El acto de atarse uno mismo debe ser un acto de comisión y no de omisión.

Es evidente que, según esta definición, la solución R del primer caso es un acto de atarse uno mismo. Por otro lado, no parece tan claro que lo sea la solución R del segundo caso. Concretamente, no parece claro que cumpla la condición 3. De hecho, la condición 3 expresa la característica distintiva del método consistente en atarse uno mismo frente a otros métodos indirectos. Estos otros métodos consisten en "una reordenación del espacio interior de la persona, sin que se establezca ningún

³ Elster (79)

mecanismo causal en el mundo externo"⁴. En opinión de Elster, estos otros métodos pueden funcionar, pero los resultados de atarse a sí mismo son siempre más "duraderos" y fiables, lo cual hace que la introducción de cambios en el entorno resulte un método privilegiado. Sin embargo, esta solución ("R2" en adelante) sí es un caso del método de atarse uno mismo. Veremos esto con algún detenimiento, ya que de esta forma se entenderá con mayor claridad qué es realmente atarse uno mismo, por qué es un método privilegiado y, lo que es más importante, por qué se habla de racionalidad imperfecta.

La solución R2 consiste en la introducción de un cambio en la estructuración de las preferencias del agente. Por lo tanto, en un sentido, consiste en un cambio "interno" y no en un cambio realizado en la situación externa. Ahora bien, un cambio en la estructuración de las preferencias puede producirse o bien mediante la operación de influencias externas⁵, o bien en ausencia de estas. Habitualmente, se habla de cambios

⁴ Elster (191), p.37.

⁵ Elster clasifica las influencias externas en cinco clases principales: seducción, persuasión, propaganda, argumentación y experiencia. La seducción se da cuando un agente que inicialmente prefería x a y y pasa a preferir y a x una vez que ha sido obligado a hacer y . La seducción, por tanto, aparece fácilmente en los casos en los que las preferencias están determinadas por los hábitos, e.d., por la experiencia pasada, en especial la experiencia reciente así como por la resistencia a la introducción de cambios. Un caso claro de este tipo de preferencias son las ligadas a la alimentación. Supongamos por ejemplo que una persona acostumbrada a condimentar la comida con mucha sal es obligado por su familia por motivos de salud a comer sin sal. En habitual que en estos casos, al cabo de cierto tiempo esa persona empieza a preferir la comida sin sal. Esto sería un caso de seducción.

La persuasión ocurre cuando un individuo que inicialmente prefiere x a y es conducido a preferir y a x mediante una sucesión de mejoras a corto plazo. Un ejemplo de persuasión podría ser el caso en el cual un individuo que inicialmente prefiere realizar sus compras en los pequeños comercios que hay al lado de su casa pasa a preferir comprar en un hipermercado, incentivado por pequeñas ventajas a corto plazo representadas habitualmente por distintos tipos de "ofertas".

La experiencia es quizá el caso más habitual de cambio de preferencias debido a influencias externas. Decimos que un cambio de preferencias es debido a la experiencia cuando un individuo que inicialmente prefiere x a y y pasa a preferir y a x tras haber realizado x . Por ejemplo, la experiencia influye en el paso de preferir comer chicle a preferir no hacerlo, una vez que ha comprobado las consecuencias que su preferencia inicial tiene para su dentadura.

Un cambio de preferencias es debido a la argumentación cuando un individuo pasa de preferir x sobre y a preferir y sobre x al haber sido convencido de la mayor deseabilidad de y mediante la defensa argumentada de las cualidades de y frente a las de x . De este modo, un individuo que, por ejemplo, prefiere la moqueta al suelo de corcho puede cambiar la ordenación de sus preferencias mediante una discusión con un amigo en la cual se evalúan los distintos méritos relativos de ambas maneras de cubrir el suelo.

Por último, tenemos la influencia de la propaganda. La propaganda se distingue de la argumentación principalmente en que el individuo que inicialmente prefiere x sobre y es llevado a preferir y

exógenos de las preferencias en el primer caso y de cambios no exógenos en el segundo.

Lo cambios no exógenos de las preferencias son los producidos en ausencia de influencias externas. Puede hablarse de dos tipos de cambios no exógenos, a saber, las preferencias temporales inconsistentes y los cambios endógenos de preferencias. Por el momento, nos ocuparemos tan sólo de los cambios endógenos, dejando para más adelante el problema de las preferencias temporales inconsistentes.

Explicar el mecanismo que subyace a los cambios endógenos de preferencias resulta extremadamente difícil, precisamente por su carácter no exógeno, e.d., por la ausencia de influencias externas cuya presencia puede explicar el cambio de las preferencias. Por otro lado, es sumamente difícil identificar ejemplos reales de cambios endógenos, puesto que habitualmente estamos sometidos a influencias externas y siempre, o al menos casi siempre, es posible explicar cualquier cambio en la estructura de las preferencias de un individuo acudiendo a tales influencias, especialmente porque no es necesario que el individuo tenga consciencia de una influencia para que esta haya sido efectiva⁴. En estos casos, la hipótesis de que estos cambios de preferencias son exógenos es más que plausible. A todas estas dificultades se añade el hecho de que no tenemos una idea muy precisa de por qué la gente prefiere lo que prefiere en los casos en los que suponemos que estas preferencias son "nativas"⁵. Este tipo de preferencias son difícilmente alterables mediante influencias externas, lo cual es un buen motivo para suponer que tampoco estas

sobre mediante una presentación no discursiva de los atributos relativos de x e y . Cualquier anuncio publicitario sirve como ejemplo.

⁴ Por ejemplo, a pesar de que es raro encontrar a alguien que reconozca la influencia de la moda a través de la propaganda, es un hecho curiosoísimo que tras unos meses de Moda de España, mucha gente sienta una afición incontenible, que antes no sentía en absoluto, por el diseño y las sacas.

⁵ Desde luego, hay casos en los que puede darse una explicación plausible ateniéndonos a las características y necesidades biológicas de los individuos. Por ejemplo, el gusto por la sal entre los habitantes de zonas muy salinosas puede explicarse por una necesidad de consumir esta sustancia a fin de retener el agua en el organismo lo más posible. Pero hay preferencias nativas que no son tan fácilmente explicables.

han tenido ningún papel en su génesis⁹. Por eso las llamamos nativas. Cuando estas cambian, habitualmente constituyen ejemplos de cambios endógenos. Sin embargo las preferencias nativas son numéricamente poco importantes. Casi todas nuestras preferencias surgen a partir de influencias externas. Cuando cambian, en muchas ocasiones lo hacen en función de influencias externas que podemos localizar con cierta facilidad. En otras ocasiones no podemos localizar esas influencias, pero tenemos motivos para pensar que han existido. Es relativamente fácil determinar si algunos tipos de influencias externas se han producido o no. Es relativamente fácil determinar si ha habido argumentación, o seducción, o experiencia. Pero es mucho más difícil saber si ha habido persuasión. Y mucho más aun saber si ha habido propaganda. Y, cuando estos mecanismos han operado, en ocasiones es difícil determinar cuál ha sido su grado de eficacia. Sin embargo, puede haber casos en los que estemos bastante seguros de que un cambio de preferencias no se debe a influencias externas. Esto sucede especialmente cuando los cambios se producen rápidamente y podemos descartar la operación de influencias externas capaces de condicionar un cambio rápido, tales como algunos tipos de experiencia y argumentación. La rapidez del cambio excluye la operación de las influencias más difíciles de localizar, como la persuasión y la propaganda, que, por su propio carácter, requieren tiempo. En estos casos, podemos hablar con bastante seguridad de cambios endógenos de preferencias.

Los cambios endógenos de las preferencias suponen un tipo de inconsistencia y, por consiguiente de irracionalidad. Ulises prefiere permanecer en el barco con sus compañeros a atender la llamada de las sirenas. Pero cuando oye su canto sus preferencias cambian. Yo prefiero no fumar más de siete u ocho cigarrillos diarios. Pero cuando me siento delante del ordenador mis preferencias cambian. El tipo de situación que representan estos ejemplos es familiar. Sin embargo, esta familiaridad no va acompañada de una facilidad para entender de qué son ejemplo estos casos. Parece poco discutible que representan un tipo de irracionalidad. Pero no está tan claro a qué es atribuible esta irracionalidad. Elster

⁹ Naturalmente, esto no es así necesariamente. Puede que mi rechazo a la seda haya surgido a partir de alguna experiencia desagradable. Pero este tipo de explicaciones no son plausibles en muchas ocasiones.

atribuye esta irracionalidad a la inconsistencia. Pero, por otro lado, trata estos casos como casos de debilidad de la voluntad. De hecho, este tipo de casos, y más concretamente el caso del tabaco, es el ejemplo típico de un problema de debilidad de la voluntad.

Ahora bien, ¿qué es lo que sucede en un caso de debilidad de la voluntad? Pensemos en el tabaco. Yo prefiero no fumar, o no fumar demasiado. Por lo tanto, decido no fumar más de cuatro cigarrillos por la mañana y tres por la tarde. Considero que el tabaco en esas cantidades no me perjudica, o al menos que el poco daño que me ocasiona queda compensado con creces por el placer que obtengo. Esta preferencia es racional: tengo en mi poder toda la información disponible sobre el tabaco, sus pros y sus contras, no me encuentro psíquica o emocionalmente alterada, mi preferencia sobreviviría a una confrontación con los hechos, no es intrínsecamente irracional y cumple todas las demás condiciones de racionalidad que en su momento consideramos exigibles⁹. Por lo tanto, mi decisión de actuar de tal modo que esta preferencia sea satisfecha es una decisión racional y llevar a cabo esta decisión es una acción racional. El caso de Ulises puede ser fácilmente analizado en la misma línea. Pero cuando me siento a escribir, o cuando Ulises oye el canto de las sirenas, la cosa cambia. El cambio que se produce implica que mis decisiones pasadas, basadas en mis preferencias, no se llevan a cabo. La cuestión está en saber por qué no se lleva a cabo mi decisión.

En un sentido, lo que sucede es que cuando llega el momento de llevar a cabo una decisión mis preferencias han cambiado, y han cambiado de tal modo que el conjunto de mis preferencias en el momento de tomar la decisión y mis preferencias en el momento de llevarla a cabo es inconsistente. Esto es verdad si suponemos que una persona siempre actúa según sus preferencias. Y, en un sentido amplio de "preferencias", esto siempre sucede. Sin embargo, esto no es decir mucho. Tenemos que saber qué tipo de preferencias se tienen en estos casos en el momento de llevar a cabo la decisión. Admitamos que han cambiado, no son las mismas de antes. Pero ¿cómo han cambiado?, ¿cómo son ahora?, concretamente, cuando me siento a escribir y mis preferencias se alteran, ¿son también racionales mis nuevas preferencias?

⁹ Ver capítulo II.

La respuesta es claramente negativa. Por una razón: mis nuevas preferencias no se han formado en un estado de ánimo adecuado. Lo mismo puede decirse de las nuevas preferencias de Ulises. Más bien, se han formado en un momento de alteración que impide la consideración serena de la información relevante. Mis nervios y el hábito del tabaco me han jugado una mala pasada, tan mala como la de las sirenas. En un sentido importante, no soy yo quien ha tomado la decisión de fumar más. En un sentido importante, la decisión me ha tomado a mí. Lo que hace irracional la conducta en estos casos no es que se base en un conjunto inconsistente de preferencias. Más bien, estas preferencias son irracionales e inconsistentes con otras preferencias que sí son racionales. Por este motivo, nuestros casos son precisamente casos de debilidad de la voluntad.

Elster parece suponer que todos los casos de cambios no exógenos de preferencias son casos de debilidad de la voluntad. Esta afirmación se defiende explícitamente para el caso de las preferencias temporales inconsistentes¹⁰, y aparece implícitamente en su tratamiento de los cambios endógenos de preferencias. Sin embargo, es precisamente en este último caso donde la identificación es más difícil de entender¹¹. Hay cosas que nos gustan y después nos dejan de gustar. Algunos de estos casos son de lo que llamamos "preferencias nativas". Y en algunas de estas ocasiones estamos razonablemente seguros de que el cambio no se ha producido por ninguna influencia externa. A mí no me gusta el dulce. Pero de vez en cuando siento incontenibles deseos de tomar algo dulce. Con toda probabilidad, esto es un cambio endógeno. Pretender que hay aquí algo irracional es poco defendible. Tampoco es cierto que mi conjunto de preferencias sea inconsistente. Simplemente, puedo preferir tomar algo dulce cuanto siento deseos de hacerlo. Esto tampoco es un caso de debilidad de la voluntad. Lo que yo he decidido es comer dulce cuando lo desee, y mi decisión se

¹⁰ Elster (79), p.67.

¹¹ De nuevo, el problema que aparece cuando se pretende discutir sobre esto es la dificultad de identificar con un cierto grado de seguridad un caso de cambio endógeno de preferencias. Por ejemplo, a los niños pequeños suelen gustarles las papillas de harina de maíz. Este gusto suele desaparecer cuando los niños se hacen un poco mayores. Puede suceder que el cambio de preferencias sea exógeno, concretamente debido a la seducción: sencillamente, cuando los niños crecen se les deja de dar papillas de harina de maíz y se les hace tomar otro tipo de alimentos. Pero podemos suponer un caso en el que un niño, que se acostaba siendo niño de estas papillas, se levanta aborreciéndolas. El caso es muy extraño, pero no es totalmente irreal.

lleva a cabo. Esto significa que un cambio endógeno de preferencias no implica ningún tipo de irracionalidad ni supone debilidad de la voluntad. Sin embargo, lo que sí es cierto es que estos casos de cambios endógenos son fácilmente transformables en casos de debilidad de la voluntad. Supongamos que yo prefiero no tomar dulce nunca. Decido no hacerlo. Pero de vez en cuando siento grandes deseos de tomar algo dulce. Cuando siento estos deseos, no puedo resistirlos. Esto sí es un caso de debilidad de la voluntad.

Diremos que algo es un caso de debilidad de la voluntad cuando consiste en la realización por parte de un agente de una acción cuyas consecuencias no son deseadas por este¹². Una acción de este tipo puede ser interpretada de dos modos. Podemos decir que el agente obra según unas preferencias irracionales. Por otro lado, podemos decir que el agente no actúa para satisfacer sus preferencias (racionales). En el primer caso, diremos que su conducta es irracional debido a la irracionalidad de sus preferencias. En el segundo, diremos que es irracional porque no está

¹² Esta definición se adecua al tratamiento clásico del problema de la debilidad de la voluntad. Por ejemplo, Aristóteles define al incontinente como aquel que obra mal sabiendo que obra mal, aquel que en sus acciones no se atiene a su razón (EN, 1145b10-14). También Platón discute el problema en estos términos.

La polémica en torno a la debilidad de la voluntad desde su planteamiento en Platón ha continuado hasta nuestros días. Esta polémica gira básicamente en torno a la propia existencia del fenómeno conocido como debilidad de la voluntad, es decir, sobre si hay casos en los que ocurre que alguien hace algo malo sabiendo que lo hace. Por un lado, la postura platónica consiste en negar que existan casos auténticos de debilidad de la voluntad, y afirma que, más bien al contrario, "era la falta de ciencia lo que obligaba a hacer una mala elección", a saber, la falta de una "cierta ciencia de la medida" que consiste en saber calcular adecuadamente los placeres y dolores sin dejarnos extraviar por las apariencias.

Por otra parte, Aristóteles mantiene que al menos hay un sentido claro en el que el fenómeno de la incontinencia existe, a saber, en que en ocasiones sabemos que algo está mal y sin embargo lo hacemos. El análisis aristotélico está más bien encaminado a descubrir en qué sentidos puede, y en cuales no puede, decirse que sabemos que lo que hacemos está mal.

Parece indiscutible que, al menos en algún sentido, sabemos que lo que hacemos está mal. No obstante, es innecesario que entráramos aquí en ella. Todo lo que nosotros necesitamos afirmar es que hay casos en los que a) tenemos unas preferencias racionales, b) estamos en ocasiones tentados a obrar en contra de estas preferencias y c) en ocasiones cedemos a esta tentación. Ni siquiera Platón niega que se den estos casos. Lo que niega es que seamos tentados por el placer. Lo que sucede, dice, es que somos cegados por las apariencias. Pero, sea lo que sea lo que hace que oremos en contra de nuestras preferencias racionales, lo cierto es que podemos de algún modo "protegermos" contra la tentación. Puede que Ulises, en el momento de oír a las sirenas, supiera de un modo u otro, que no debía irse con ellas, o puede que su canto engañoso le hiciera olvidar por completo todo lo que sabía acerca de las consecuencias de atender a sus llamadas. Pero en cualquier caso antes de oírlo, sabía a) que no debía irse, b) que si oírlo, estaría tentado a obrar en contra de sus deseos (actuales) y que c) no podría resistir la tentación. Esto es suficiente para que pudiera curarse en salud y pedir que le ataran al mástil y no le soltaran pasara lo que pasara.

dirigida a la satisfacción de sus preferencias.

Es difícil decidir cuál de las dos interpretaciones es correcta. Cuando yo enciendo un cigarrillo que no debo encender, es indudable que, en un sentido, lo hago porque prefiero hacerlo. En este sentido, toda acción voluntaria¹³ es una acción según las preferencias¹⁴. Por otro lado, parece dudoso que yo "realmente" prefiera encender el cigarrillo en cuestión. Más bien parece que al hacerlo obro de un modo irreflexivo y que, si en el momento de ir a encenderlo, alguien me detuviera y me preguntara "¿seguro que quieres encenderlo?" yo diría "no". Sin embargo, esto no es exacto. Mi respuesta inmediata sería "sí". Si mi interlocutor prosiguiera diciendo "pero considera tal y tal cosa", entonces pueden suceder cosas distintas. Es probable que cambiara de opinión (posibilidad A). Y también puede que en algunas circunstancias contestara "sí ya lo sé, pero ahora mismo me da igual" (posibilidad B) o incluso "sí ya lo sé, pero la tentación es demasiado fuerte para mí" (posibilidad C). En los casos A y B, parece que es más exacto decir que yo actúo según mis preferencias, que estas son irracionales, y que las distintas respuestas ofrecidas por mí son otras tantas razones para pensar que mis preferencias son irracionales. Sólo en el caso de que ocurra C parece que la segunda interpretación es adecuada. Parece por consiguiente que ambas interpretaciones pueden resultar correctas según los casos¹⁵.

Si yo quiero comer dulce cuando lo desee, la cuestión del uso de estrategias no se plantea. Esta surge en los casos en los que yo quiero algo y sé que, cuando llegue el momento, no voy a ser capaz de mantener

¹³ Hablamos aquí de acciones "voluntarias" e "involuntarias" en el sentido aristotélico, según el cual una acción es involuntaria o bien cuando es forzada, e.d., "aquellas cuya causa está fuera del agente y en las que este no tiene parte alguna", o bien cuando es debida a "la ignorancia concreta de las circunstancias y el objeto de la acción", aunque a las acciones de este último tipo prefiera llamarlas "no voluntarias" (Aristóteles EN, 1110a-1111a).

¹⁴ Puede objetársele a esta interpretación que al hacer equivalentes "acción voluntaria" y "acción acorde con las preferencias", y dada la definición de "acción racional" como "acción maximizadora de la utilidad", se seguiría que toda acción voluntaria es una acción racional, lo cual no parece verdad salvo que lo sea por definición. Sin embargo, esta objeción no es válida, puesto que no se define acción racional sólo como "acción maximizadora de la utilidad", sino que se exige que la utilidad esté definida sobre el conjunto racional de las preferencias de un agente.

¹⁵ Nótese que el motivo por el que la segunda interpretación parece la más adecuada en el caso C es porque esta se parece extraordinariamente (veremos más adelante en qué sentido) a las acciones involuntarias.

mi decisión. Esto significa que el problema no está en el cambio endógeno de las preferencias, sino más bien en la debilidad de la voluntad. Es decir, que algo sea un caso de debilidad de la voluntad es precisamente lo que hace válido y necesario el uso de estrategias indirectas. Ahora bien, lo que hace posible hablar de debilidad de la voluntad no es la presencia de dos preferencias contrapuestas, sino el hecho de que una de estas sea una preferencia racional mientras que la otra es una preferencia irracional. Esto es lo que nos permite identificar una de ellas como "nuestra" auténtica preferencia y a la otra como algo que nos sobreviene¹⁴. Por eso queremos ofrecer resistencia a estas últimas. En realidad, la decisión que tomamos, basada en nuestras preferencias racionales, es precisamente la decisión de resistir a las preferencias irracionales cuando aparecen. Y por eso, cuando dudamos de nuestra fuerza, de la fuerza de nuestra voluntad, acudimos a estrategias. Todas las mañanas, cuando me siento a escribir, cojo cuatro cigarrillos y le doy a otra persona el resto del paquete, pidiéndole que no me lo devuelva hasta después de comer. Los únicos motivos que tengo para hacer esto son que, por un lado, me identifico con mi decisión de no fumar más de siete u ocho cigarrillos diarios, e.d., que considero que esta es la decisión basada en mis preferencias racionales, formadas cuando no estoy sometida a circunstancias perturbadoras y, por otro lado, que sé perfectamente que mi fuerza de voluntad no es suficiente. Por eso transformo la situación de tal modo que lo que yo haga ya no dependerá de mi voluntad, hacia la cual tengo motivos sobrados de desconfianza.

Sin embargo, hay una razón para suponer que todo cambio endógeno de las preferencias implica un problema de debilidad de la voluntad. Antes de analizar esta razón conviene recordar una distinción terminológica realizada en el capítulo I. Llamábamos "preferencias primarias" a las preferencias ligadas a los placeres y dolores puramente físicos tomados por separado. Como se recordará, decíamos que hay un sentido en el que estas preferencias no podían ser irracionales. Pues bien, en este mismo sentido el cambio endógeno de las preferencias primarias tampoco puede ser irracional. Normalmente, no me gusta el dulce, pero a veces siento un

¹⁴ Habitualmente, no soy partidaria de este tipo de expresión un tanto esquizofrénico, pero creo que en esta ocasión, y entendido con las debidas precauciones, resulta extraordinariamente útil.

gran deseo de algo dulce. Ninguna de las dos preferencias es irracional. El conjunto de mis preferencias primarias a través del tiempo puede contener un par de preferencias opuestas sin que por ello pueda hablarse de inconsistencia. Esto sucede porque las preferencias primarias son necesariamente temporales. El conjunto de mis preferencias primarias a través del tiempo contendría en nuestro ejemplo el par "no me gusta el dulce en t_1 ", "me gusta el dulce en t_2 ". La referencia a distintos momentos hace que el conjunto de mis preferencias no sea inconsistente. Por eso, el cambio endógeno de este tipo de preferencias no resulta problemático.

El cambio de preferencias sólo es problemático en el caso de las preferencias secundarias. Estas involucran la elección entre distintas preferencias primarias. En nuestro ejemplo, una posible preferencia secundaria sería "prefiero no tomar dulce cuando no me apetece y tomarlo cuando siento deseos de hacerlo". Esta preferencia secundaria consiste simplemente en la preferencia de obrar según nuestras preferencias primarias. Este tipo de preferencias pueden hacer referencia al tiempo. Es decir, la preferencia anterior está formulada atemporalmente, pero sería equivalente a un conjunto de preferencias temporales del tipo "prefiero no comer dulce en t_1 ", "prefiero tomar dulce en t_2 ". Aunque en esta formulación aparentemente las preferencias primarias y secundarias se identifican, son bien distintas. Las preferencias secundarias significan "En el momento t_1 (e.d., en el momento en el que no me apetece tomar dulce), elijo no tomar dulce" y "En el momento t_2 (e.d., cuando me apetece tomar dulce), elijo tomar dulce" respectivamente. Por el contrario, las preferencias primarias no involucran necesariamente una decisión o una elección, en el sentido de que la decisión o la elección no se siguen necesariamente de una ordenación de estas preferencias, cosa que sí sucede a partir de una ordenación de preferencias secundarias, a las que, por este motivo, nos referimos simplemente como "preferencias".

En este ejemplo, nuestras preferencias secundarias no cambian. Pero pueden cambiar. Por ejemplo, yo ahora puedo preferir no tomar dulce (e.d., no tomar dulce nunca), y en un momento posterior puedo preferir tomar dulce. El cambio es posible porque estas preferencias pueden ser (aunque no son necesariamente, como se muestra en el caso anterior) temporalmente neutrales, e.d., pueden no estar ligados a un momento determinado. Esto también hace posible hablar de inconsistencia en un conjunto

de preferencias secundarias. Como ya hemos dicho, lo que hace posible hablar de estrategias indirectas destinadas a resistir los cambios endógenos de preferencias es que una de las preferencias sea una preferencia irracional, cosa que al mismo tiempo nos permite identificar este problema como un problema de debilidad de la voluntad. Ahora bien, supongamos un caso en el que se produce un cambio de preferencias y que no es sin embargo un caso de debilidad de la voluntad. Yo he decidido no tomar nunca dulce. Como dijimos en el capítulo II, esta preferencia es intrínsecamente irracional si consiste en preferir lo peor. Veamos qué sucede en el momento en que me gusta el dulce. En ese momento, mi preferencia de no tomar dulce es irracional a menos que yo pueda justificarla haciendo ver que mi preferencia es en realidad de lo mejor, por ejemplo argumentando que si tomo dulce engordaré y eso es para mí lo peor. Supongamos que mi preferencia cambia, de tal modo que prefiero tomar dulce. ¿Que puede haber pasado? El cambio puede deberse a tras cosas:

-Puede que el cambio en las preferencias secundarias responda a un cambio en las preferencias primarias, por ejemplo, a que empiece a gustarme (o a no disgustarme) estar un poco gorda. Este caso no resulta problemático, puesto que desaparece mi preferencia al desaparecer la razón que la justificaba.

-Puede que mi nueva preferencia surja de una debilidad de la voluntad. Este es el tipo de caso que hemos analizado hasta ahora.

-Por último, puede que piense ahora que un poco de dulce, dada mi constitución y mi actividad, no me hará daño.

En este último caso, el cambio de las preferencias no implica un caso de debilidad de la voluntad. Tampoco es irracional la nueva preferencia¹⁷. Estas dos cosas son suficientes para que no pueda plantearse el uso de estrategias indirectas destinadas a resistir el cambio de las preferencias. La cuestión es si en este caso es posible que el cambio sea un cambio endógeno. La respuesta parece ser evidentemente negativa. El cambio ha debido producirse a causa de algún tipo de influencia externa. Puede que un médico amigo mío me haya convencido por medio de argumentos.

¹⁷ En todo caso, puede argumentarse que era la antigua preferencia la que era irracional, en el sentido de que estaba basada en una información defectuosa sobre los efectos del dulce, por lo cual no habría sobrevivido (así bien, no ha sobrevivido) a una confrontación con los hechos y la información relevante.

O puede que haya aprendido por experiencia que un poco de dulce no es malo. Y es necesario que así sea, puesto que el cambio se ha debido a que las razones que justificaban mi anterior preferencia han desaparecido. Por eso, una vez subrayada la distinción entre preferencias primarias y preferencias secundarias, podemos afirmar que todo cambio endógeno de preferencias (secundarias) es realmente un caso de debilidad de la voluntad, es decir, un caso en el que las decisiones no se mantienen a pesar de que no se han alterado las razones que justifican nuestra decisión.

Hasta ahora hemos analizado el cambio de las preferencias¹⁶. Por un lado, hemos visto las influencias externas que pueden dar lugar a cambios exógenos y, por otro, hemos analizado los cambios endógenos interpretándolos como un problema de debilidad de la voluntad. Los cambios no exógenos (e.d., los cambios endógenos y las preferencias temporales inconsistentes, a las cuales aún no hemos dedicado nuestra atención) son cambios no deliberados. Puesto que se deben a la debilidad de la voluntad, son cambios indeseados, cambios contra los que el agente desea protegerse. Si yo prefiero no fumar más de ocho cigarrillos diarios y, basándome en esta preferencia, tomo una decisión, deseo que esta se lleve a cabo. Pero para poder satisfacer esta preferencia puedo tener que protegerme de diversos factores que pueden intervenir impidiendo o dificultando la realización de mi decisión. Algunos de estos factores son influencias externas. Pero otros no lo son. Más bien, consisten en un cambio endógeno de mis preferencias. En estos casos, de alguna forma lo que necesito es protección contra mí misma o, dicho con más exactitud, contra mi propia irracionalidad. Los métodos indirectos de racionalidad son los mecanismos con los que contamos para procurarnos esta protección. Y entre estos mecanismos ocupa un lugar privilegiado el método consistente en atarse uno mismo¹⁷.

Atarse uno mismo es algo que puede hacerse de dos maneras. Estas dos maneras están ligadas a los dos procesos que determinan el que un

¹⁶ Salvo que se indique expresamente lo contrario, siempre que hablemos de preferencias o cambio de preferencias nos referiremos a las preferencias secundarias.

¹⁷ Atarse uno mismo no sólo es un remedio para resistir los cambios endógenos de las preferencias. Esto sucede en el caso del tabaco. Pero en el caso de Ulises, por ejemplo, atarse uno mismo resulta una protección contra determinadas influencias externas, es decir, contra determinados cambios exógenos de las preferencias.

agente actúe de determinada manera en determinada situación

"Para explicar por qué una persona en una situación dada se comporta de una manera y no de otra podemos ver su acción como el resultado de dos procesos de filtración sucesivos. El primero tiene como efecto limitar el conjunto de acciones abstractamente posibles al conjunto factible, e.d., al conjunto de acciones que satisfacen simultáneamente un número de constricciones físicas, técnicas, económicas y político-legales. El segundo tiene el efecto de determinar un miembro del conjunto factible como el miembro que va a ser llevado a cabo"²⁰.

El acto de atarse uno mismo puede dirigirse a la intervención en uno de estos dos procesos. En primer lugar, atarse uno mismo puede tener como efecto la modificación deliberada del conjunto de alternativas factibles. Este modo de atarse uno mismo es más sencillo y probablemente el que tiene más garantías de éxito, motivo por el que sin duda es el paradigma de atarse uno mismo. Ulises se ata, eliminando así la posibilidad física de acudir a la llamada de las sirenas. Yo le doy el tabaco a otra persona, o lo alejo de mi alcance de algún otro modo para eliminar la posibilidad física de disponer de tabaco. A pesar de sus indudables ventajas, este procedimiento no siempre es el mejor. A veces, manipular el conjunto factible de posibilidades resulta extremadamente costoso. Puedo salir al campo sin tabaco para evitar fumar. Pero no puedo llevarme el ordenador, ni puedo quedarme allí a vivir. La solución es buena en algunos momentos, pero no es una solución definitiva. Ulises puede pedir que le aten porque el canto de las sirenas sólo se oye en un determinado lugar. Pero si el canto de las sirenas fuera continuo, atarse físicamente no sería una buena solución. Le impediría, desde luego, acudir a su llamada. Pero también le impediría hacer cualquier otra cosa. A veces, este procedimiento no sólo resulta costoso, sino tremendamente ineficaz. Puedo eliminar con relativa facilidad algunas posibilidades de conseguir tabaco. Pero no puedo eliminarlas todas. Puedo salir sin dinero. Pero puedo pedirlo. Puedo pedir a los conocidos que no me lo den. Pero hay desconocidos. Conseguir que este método funcione de manera eficaz elevaría el coste a niveles absurdos.

Por todo esto, en ocasiones la manera más eficaz y definitiva de atarse uno mismo consiste en la manipulación, no del conjunto factible de

²⁰ Elster (79), p.77.

elecciones, sino del mecanismo por el cual un miembro de este conjunto resulta elegido. Pero manipular este mecanismo supone manipular deliberadamente las propias preferencias, puesto que en último término son las preferencias del agente las que determinan que uno de los miembros del conjunto de acciones factibles, y no otro, sea escogido²¹.

En resumen, la resistencia a los cambios endógenos puede ejercerse tanto por el desarrollo de la voluntad como por una manipulación de la situación encaminada a conseguir que la cantidad de voluntad necesaria sea lo más pequeña posible.

Podemos pasar a ocuparnos ahora de la *segunda* posibilidad de manipulación de las preferencias, consistente en la introducción deliberada de un cambio en la estructuración de las preferencias. En esta ocasión no se trata de impedir un cambio, sino de introducirlo. En la mayoría de las ocasiones, introducir un cambio de este tipo supone la exposición deliberada a las influencias externas. Supongamos por ejemplo que yo siempre prefiero comer con sal pero que, por algún motivo, deseo cambiar esa preferencia. Puesto que sé que mi preferencia por la sal depende del hábito, puedo pedirle a quien haga la comida que no ponga sal. Es decir, puedo someterme a los efectos de un tipo de influencia externa, concretamente a la que llamamos seducción. O supongamos que siempre prefiero viajar en tren pero que pienso que me iría mejor si mi preferencia se alterara a favor del avión. Entre otras cosas, puedo apuntarme a un curso impartido por unas líneas aéreas en el que, mediante la utilización la argumentación y la propaganda, mis preferencias se alterarán. El objeto de estas manipulaciones es introducir un cambio en mis preferencias. Pero para conseguir ese cambio interno realizo una serie de acciones destinadas a

²¹ Elster señala que hay dos modos de manipular las propias preferencias. En unas ocasiones, esta manipulación consiste en intentar deliberadamente introducir un cambio en la estructuración de las preferencias, y, en otras, en resistir deliberadamente un cambio endógeno. A esta segunda modalidad de auto-manipulación es a la que Elster dedica toda su atención. Puesto que los cambios endógenos de las preferencias suponen un problema de debilidad de la voluntad, la resistencia a estos cambios supone o bien un reforzamiento de la voluntad o bien una resistencia a la formación de hábitos que, una vez desarrollados, son demasiado fuertes. El tabaco, o cualquier otra droga, son ejemplos paradigmáticos de este tipo de hábitos. Esto supone en muchos casos la resistencia a exponerse a determinadas influencias externas. Por ejemplo, en el caso del tabaco, la resistencia a la adquisición del hábito puede requerir el no exponerse a la publicidad. Esto sucede porque, aunque una vez que el hábito está formado se presenta el fenómeno de la debilidad de la voluntad y de los consiguientes cambios endógenos, el cambio de preferencias que supone la formación del hábito puede ser (suele ser) de carácter exógeno.

establecer un proceso causal en el mundo externo. Esto significa que mis acciones son casos del método indirecto de atarse uno mismo, puesto que cumplen todas las características de este método, incluida la condición 3.

Una vez que hemos analizado las distintas maneras de atarse uno mismo, podemos entender mejor por que es en ocasiones necesario el uso de esta estrategia indirecta. Un agente racional actúa siempre de tal modo que maximiza con su conducta la utilidad esperada. Es decir, un agente perfectamente racional, que actuara siempre movido por este motivo, no necesitaría usar este tipo de estrategias. Pero nosotros no somos agentes perfectamente racionales. Pensemos en Ulises

"Ulises no era enteramente racional, porque una criatura racional no habría tenido que recurrir a esa artimaña; tampoco era simplemente el vehículo irracional de sus preferencias y deseos cambiantes, porque era capaz de conseguir por medios indirectos el mismo fin que un agente racional hubiera podido alcanzar de modo directo."²²

Al hablar así, Elster supone que el motivo por el que necesitamos acudir a métodos indirectos es nuestra racionalidad imperfecta. El principal motivo por el que no somos perfectamente racionales es que algunas de nuestras preferencias son irracionales. Es irracional la preferencia de Ulises cuando oye a las sirenas y es irracional mi preferencia por fumar cuando me siento a escribir²³.

Sin duda, Elster tiene razón al afirmar que no somos perfectamente racionales y que, por ello, necesitamos acudir a métodos indirectos. Sin embargo, no parece claro que este sea el único motivo por el que necesitamos estos métodos. Concretamente, no parece claro que las soluciones R1 y R2 a los ejemplos con los que iniciamos esta sección y que, como hemos

²² Elster (79), p.36.

²³ Como hemos visto, Elster atribuye esto a la debilidad de la voluntad. Puesto que nuestra voluntad no siempre es lo suficientemente fuerte como para que hagamos lo más racional, necesitamos algún tipo de estrategia. Puede verse entonces claramente el sentido de la condición 3. Puesto que el problema reside precisamente en nuestra voluntad, no podemos solucionar el problema confiando en ella. Más bien al contrario; la solución indirecta ha de consistir en establecer determinados mecanismos en el mundo externo, independientes de nuestra voluntad. Nótese que esto es así incluso en los casos en los que la estrategia está dirigida a modificar nuestras propias preferencias. Para modificarlas, acudimos a someternos a determinadas influencias externas. Modificarlas de otro modo supondría precisamente estar en poder de la fuerza de voluntad que no tenemos y que pretendemos suplir mediante el uso de estrategias.

visto, sean soluciones indirectas, e.d., sean necesarias por un problema achacable a la debilidad de la voluntad. En el caso 1, como se recordará, el problema era el siguiente. Lo mejor para mí es que el taxista me lleve a casa, pero sólo me llevará si cree que le pagaré. De modo que lo mejor para mí es prometerle que lo haré. Pero una vez que llego a casa lo mejor para mí es no pagar. En esta situación, la acción racional consiste en prometerle al taxista que le pagaré y, una vez que estoy en casa, lo racional es no pagarle. El problema empieza cuando suponemos que somos transparentes. El taxista sabe que yo soy un agente racional, y que siempre hago lo mejor para mí. Pero entonces sabe que no le pagaré. De modo que no me llevará y el resultado será para mí el peor posible. En este caso, la debilidad de la voluntad no desempeña ningún papel. Supongamos que yo decido pagarle y le prometo que lo haré. Cuando llego a casa, la cosa cambia. Pero este cambio no es atribuible a la debilidad de la voluntad. Lo que sucede no es que, debido a la debilidad de mi voluntad, cuando llega el momento de pagar sea incapaz de llevar a cabo mi decisión. Más bien, lo que sucede es que mi propia racionalidad me impide pagarle. Lo que diferencia este caso del caso del tabaco es fundamentalmente lo siguiente. En el caso del tabaco, mi preferencia por fumar, y por tanto, por no llevar a cabo la decisión de no fumar, cuando me siento a escribir es una preferencia irracional. Pero en este otro caso mi preferencia por no pagar, e.d., por no llevar a cabo la decisión de pagar, es racional. En el caso del tabaco, es la debilidad de mi voluntad la que me juega una mala pasada. Pero en este otro caso es mi propia racionalidad quien lo hace. Si necesitamos aquí un método indirecto no es para defendernos de nuestra propia irracionalidad, sino más bien de nuestra propia racionalidad.

Algo parecido sucede en el caso 2. Si actúo según mis preferencias racionales, a saber, mi interés por mí y por mis hijos, el resultado será pésimo. Yo sé que actuar según estas preferencias tendrá este resultado. La única salida posible es tomar la droga que cambia mis preferencias. Comparemos esto con el caso del tabaco y supongamos que dispongo de una droga que evita que al sentarme a escribir, prefiera fumar. En este caso, el motivo por el que tomo la droga es que no quiero satisfacer mi preferencia por fumar y el mejor modo de evitarlo es no tener esta preferencia. Pero en el otro caso sucede algo bien distinto. El motivo por el que

tomo la droga es que *quiero que se satisfagan mis antiguas preferencias y el mejor modo de conseguirlo es no tenerlas*. La preferencia que quiero evitar tomando la droga contra el tabaco es una preferencia irracional que no quiero satisfacer. Pero en el otro caso, mi preferencia por satisfacer mis intereses y los de mis hijos es una preferencia racional que quiero satisfacer.

Sin embargo, los casos 1 y 2 se parecen bastante a los casos analizados por Elster. En todos ellos, lo racional es tomar una determinada decisión y llevarla a cabo. Cuando estoy en medio de la calle sin dinero, la ordenación de los distintos resultados alternativos según mis preferencias es el siguiente. En primer lugar, prefiero que el taxista me lleve a casa y no tener que pagar, en segundo lugar prefiero que me lleve y pagar y, por último está la alternativa de que el taxista no me lleve. Pero, supuesto que somos transparentes, la alternativa que yo prefiero en primer lugar simplemente es inalcanzable. Lo mejor que puedo hacer es intentar conseguir el segundo resultado. Cualquier teoría razonable acerca de la racionalidad nos diría que tenemos que hacer lo posible para conseguir este resultado que, dadas las circunstancias, es el mejor posible.

El problema es que este resultado, siendo las cosas como son y siendo yo como soy, es también inalcanzable. Sé que cuando llegue el momento, tendré la tentación irresistible de no pagar. Es decir, tanto en este caso como en los planteados por Elster, el problema es que una decisión tomada en un momento t_1 no se mantiene en el momento t_2 . Esto quiere decir que, en un sentido, todos los casos son de debilidad de la voluntad, a saber, en el sentido de que la voluntad falla cuando hay que llevar a cabo una decisión tomada. Pero en otro sentido, el ejemplo del taxista no presenta un problema de debilidad de la voluntad. Habitualmente, entendemos que la debilidad de la voluntad obedece a la imposibilidad de determinar las acciones por la razón. En este sentido, las tentaciones que aparecen en los casos de auténtica debilidad de la voluntad son tentaciones de obrar irracionalmente. Sin embargo, en el caso del taxista la tentación es una tentación racional. Por eso puede ser confuso hablar indiscriminadamente de debilidad de la voluntad.

Por tanto, podemos concluir que las estratagemas indirectas no sólo son soluciones para los casos de debilidad de la voluntad, sino, en

general, para todos aquellos casos en los que

a) exista un resultado que sea el racionalmente deseable (e.d., el mejor resultado posible) y

b) las circunstancias sean tales que este resultado sea inalcanzable sin la utilización de tales métodos. Dicho de un modo más preciso, las estrategias indirectas son aplicables en todos aquellos casos en los que a) lo mejor (e.d., lo racionalmente deseable) es tomar en un momento t_1 una decisión y llevarla a cabo en un momento t_2 y b) cuando llega el momento t_2 el agente no es capaz de llevar a cabo la decisión tomada en t_1 .

Los casos 1 y 2 con los que comenzamos este capítulo plantean situaciones en las que es necesario utilizar estrategias indirectas. Hasta ahora hemos analizado las soluciones R a estos casos y hemos visto que son soluciones indirectas, concretamente, son ejemplos del método indirecto de atarse uno mismo. Podemos volver ahora nuestra atención a las soluciones I a estos mismos casos. Para ello, veamos en qué se diferencian estas soluciones de las anteriores.

Llamamos R (racionales) a las soluciones que hemos analizado porque son soluciones internas a la teoría de la racionalidad. Es decir, mediante su utilización el agente consigue el resultado racionalmente deseable actuando racionalmente. En el caso 1, la solución R consiste en que le deje en prenda al taxista algo lo suficientemente valioso como para que lo racional sea volver a pagarle. En el caso 2, la solución R consiste en tomar la droga(2), cuyos efectos son los de alterar el orden de mis preferencias. Cuando vuelvo a pagar al taxista y a recuperar mi prenda, o cuando obro según mis preferencias alteradas por la droga (2) me estoy comportando de un modo perfectamente racional²⁴.

²⁴ Puede pensarse que en el caso 2 esto es más que discutible, puesto que puede dudarse que las preferencias que se adquieren al tomar la droga sean preferencias racionales. Concretamente, puede dudarse que tales preferencias puedan sobrevivir a una confrontación con los hechos relevantes: estando el agente en un estado de ánimo no perturbado. Más bien puede parecer que el hecho de que estas preferencias sean producidas por una droga y que sólo se mantengan mientras duran los efectos de estas, es suficiente para asegurar que el sujeto no está en un estado de ánimo adecuado. Parece que el uso de la droga altera por definición el estado de ánimo del agente. En un sentido, esto es cierto. Pero no lo es totalmente precisamente porque la decisión de tomar la droga es deliberada y el sujeto la realiza conociendo todos los hechos relevantes, entre los cuales se cuentan los efectos de la droga, en un estado de ánimo sereno y precisamente porque se quieren adquirir tales preferencias. Dicho de otro modo, si la preferencia por tomar la droga y cambiar las preferencias es perfectamente racional, lo cual es suficiente

Por el contrario, las soluciones I (irracionales) no son en ese sentido internas, puesto que en ellas el agente actúa de un modo que la teoría considera irracional. En el caso 1, yo me transformo en un individuo que cumple lo prometido con independencia de que hacerlo sea o no lo mejor para él. Es decir, me transformo en un individuo que no siempre obra racionalmente. Cuando vuelvo a pagar al taxista estoy haciendo lo peor para mí y, por ello, estoy actuando de un modo irracional. En el caso dos, la droga(1) tiene como efecto volverme un individuo irracional, en el sentido de que me convierto en un individuo que prefiere lo peor.

Podemos generalizar esto del modo siguiente. Según la condición 1 que una estratagema debe cumplir para ser un caso de atarse uno mismo²³, atarse a uno mismo es llevar a cabo una cierta decisión en el momento t_1 , a fin de incrementar la probabilidad de llevar a cabo otra decisión en t_2 , de tal modo que este incremento de la probabilidad de la acción posterior sea el motivo de la primera. Pues bien, diremos que un acto de atarse uno mismo es una solución R cuando la acción que se ha de realizar en t_2 y cuya probabilidad de realización se desea incrementar, es una acción racional y diremos que un acto de atarse uno mismo es una solución I cuando la acción a realizar en t_2 es una acción irracional.

para afirmar que las preferencias alteradas no son irracionales en ese sentido. Y puesto que tampoco son irracionales en ningún otro sentido (por ejemplo, no son intrínsecamente irracionales), podemos afirmar que son preferencias racionales.

²³ Véase que esta condición forma parte de la definición de estratagema indirecta en general. Lo que distingue al método de atarse a uno mismo de otras formas de racionalidad indirecta es el cumplimiento de la condición 3.

II. *Qué es lo que nos pide nuestra teoría*

Las soluciones I plantean a nuestra teoría de la racionalidad un problema que no plantean las soluciones R, al consistir en la realización de acciones que la propia teoría considera irracionales. Debemos por tanto preguntarnos qué sucede cuando en una determinada situación las únicas soluciones posibles son las soluciones I. Sin duda, este tipo de situaciones pueden darse. Por ejemplo, en el caso 1 puede suceder que yo no tenga nada suficientemente valioso para dejar en prenda o, en el caso 2, puede que yo sólo disponga de la droga(1). En cualquier caso, lo importante es saber qué ocurriría en este tipo de situaciones. Es decir, saber que debería hacer un agente racional que se encontrara en una de esas situaciones.

Al igual que nuestra teoría nos aconseja adoptar las soluciones R en los casos en los que es necesaria su utilización, podemos afirmar que igualmente nos aconseja adoptar las soluciones I, y por los mismos motivos. Es decir, parece que cualquier teoría aceptable acerca de la racionalidad nos pediría que, en el caso 1, nos volviéramos personas que mantienen su palabra y, en el caso 2, que tomáramos la droga(1), supuesto que esta es la única manera, o al menos la mejor manera, de maximizar nuestro nivel de utilidad. Sin embargo, esto tiene un aire paradójico, puesto que significa que la teoría de la racionalidad nos pediría que, en determinadas circunstancias, actuáramos de un modo irracional.

Para tratar esta paradoja, podemos empezar por preguntar qué es lo que pide la teoría. Para esto podemos acudir a una distinción muy útil establecida por Parfit

"Podemos describir cualquier teoría diciendo qué es lo que nos pide que intentemos conseguir. Según todas las teorías morales, debemos intentar actuar moralmente. Según todas las teorías acerca de la racionalidad, debemos intentar actuar racionalmente. Digamos que estos son los objetivos *formales*. Diferentes teorías morales y diferentes teorías acerca de la racionalidad nos proponen distintos objetivos *sustantivos*."²⁴

Todas las teorías acerca de la racionalidad nos plantean el mismo objetivo formal. Lo que las distingue son los objetivos sustantivos. Estos son

²⁴ Parfit (1961), p.3.

la especificación de lo que se entiende dentro de la teoría como acción racional.

Una vez que se ha realizado esta distinción, es fácil ver que la paradoja planteada no es algo exclusivo de nuestra teoría. Más bien, esta se planteará siempre que el mejor modo de alcanzar el objetivo sustantivo que nos da la teoría sea no intentar alcanzarlo. Y esto puede suceder con cualquier objetivo sustantivo, al menos en tanto que este es distinto del objetivo formal.

Todas las teorías deben tener un objetivo sustantivo. Esto no es sólo lo que las distingue de las demás, sino que es lo que constituye propiamente la teoría. Una teoría acerca de la racionalidad es una teoría vacía a menos que nos diga qué es lo racional. Esto significa que el objetivo formal no puede ser el objetivo sustantivo, o, al menos, que no puede ser *todo* el objetivo sustantivo. Sin embargo, es posible que el objetivo formal sea parte del objetivo sustantivo. Pensemos por ejemplo en una teoría sobre la racionalidad distinta de la nuestra, a la que Parfit da el nombre de Teoría de la lista objetiva²⁷. Según esta teoría, a la que nos referiremos como (O) ciertas cosas son buenas o malas para nosotros, con independencia de que nosotros queramos las cosas buenas o de que deseemos evitar las cosas malas. Esta teoría puede tener distintas versiones que se diferencian unas de otras por el contenido de la lista de las cosas buenas y de las cosas malas. Así por ejemplo, entre las cosas que suelen ser consideradas buenas se encuentran el conocimiento, el desarrollo de las propias capacidades o la contemplación de la belleza. Una versión de (O) no tiene por qué contener una lista con una sólo cosa buena. Es posible, incluso habitual, que varias cosas como las puestas como ejemplo, o incluso todas ellas, formen parte de la lista. Puede suceder que una lista que contenga cosas del tipo señalado sea todo el objetivo sustantivo. Pero es también posible que la teoría nos proponga como parte del objetivo sustantivo que obremos siempre racionalmente.

²⁷ Según Parfit, esto es una versión distinta de la Teoría del interés propio (S), de la cual nuestra teoría es también una versión. El objetivo sustantivo que S nos da a cada uno de nosotros es que nuestra vida sea para nosotros lo mejor posible. Las distintas versiones de S se diferencian unas de otras simplemente en lo que cada una de ellas considera que hace que nuestra vida sea lo mejor posible, es decir, en que cada una de ellas se apoya en una teoría distinta acerca del auto-interés (ver Parfit, 86 p. 3-4). Sin embargo, nuestra teoría no es ninguna de las versiones de S analizadas por Parfit. En su momento veremos si T es una versión de S y cuáles son las teorías alternativas a S.

e.d., que considere que una de las cosas buenas para nosotros es obrar racionalmente, sean cuales sean los resultados de nuestro obrar racional. Para una teoría de este tipo el objetivo formal sería una parte del objetivo sustantivo.

Que el objetivo formal sea o no parte del objetivo sustantivo resulta de gran importancia para resolver nuestra paradoja. El motivo es el siguiente. Supongamos que la mejor manera de conseguir las cosas buenas para nosotros es no perseguirlas. En algunas versiones de (O) es bastante probable que se den situaciones de este tipo, por ejemplo si una de las cosas buenas que aparecen en la lista es la contemplación de la belleza. Entonces, si el objetivo formal no es una parte del objetivo sustantivo, la teoría podría afirmar que en ese caso deberíamos dejar de perseguir la contemplación de la belleza. Es decir, podría pedirnos que actuáramos de un modo irracional. Pero si obrar racionalmente es parte del objetivo sustantivo de la teoría, entonces está menos claro lo que debería hacerse en ese caso. La respuesta dependería de cosas tales como el distinto grado de bondad de los miembros de la lista, e.d., de si la contemplación de la belleza es igual de bueno o más o menos bueno que el obrar racionalmente. Un conflicto de este tipo se solventaría del mismo modo que cualquier conflicto entre dos miembros de la lista, e.d., sería un conflicto del mismo tipo que el que podría surgir en una versión de (O) cuya lista contuviera la contemplación de la belleza y la búsqueda del conocimiento intelectual en los casos en los que no pueden conseguirse ambas cosas al mismo tiempo.

Nuestra teoría (T), al igual que todas las otras teorías alternativas posibles, nos pide que actuemos de forma racional. Pero lo específico de T es el objetivo sustantivo. Este objetivo²⁸ es el expuesto en capítulo III. Según T, una acción es racional si maximiza la utilidad esperada, definida sobre las preferencias racionales del agente. Y un agente es racional si realiza acciones racionales. Parece entonces que actuar racionalmente no es parte del objetivo que T nos plantea. Por supuesto, habitualmente la mejor forma de alcanzar el objetivo que nos da una teoría es intentar alcanzarlo. Y, en tanto que intentar alcanzar el objetivo

²⁸ En adelante, y salvo que se especifique lo contrario, por 'objetivo' entenderemos siempre 'objetivo sustantivo'.

sustantivo de una teoría de la racionalidad es actuar racionalmente, la teoría nos pide que actuemos así. Pero esto significa que actuar racionalmente es sólo un medio. Por lo tanto, en los casos en los que actuar racionalmente hace que los objetivos sustantivos de la teoría no sean alcanzados, la teoría puede pedirnos que obremos irracionalmente.

Que nuestra teoría de lugar a estas paradojas puede parecer una objeción contra ella. Esta objeción puede concretarse diciendo que T es una teoría auto-frustrante. Podemos distinguir dos modos en los que una teoría puede ser auto-frustrante a nivel individual:

-diremos que una teoría es indirectamente auto-frustrante a nivel individual "cuando es verdad que, si alguien trata de alcanzar los objetivos planteados por la teoría entonces esos objetivos serán en conjunto peor alcanzados"²⁹

- por otra parte, diremos que una teoría es directamente auto-frustrante a nivel individual "cuando es cierto que, si alguien sigue con éxito la teoría, entonces y por ese motivo hará que los objetivos planteados por la teoría sean peor alcanzados de lo que lo hubieran sido si él no hubiera seguido la teoría con éxito"³⁰

Una teoría como la nuestra no sólo no es directamente auto-frustrante, sino que no es posible que lo sea. Que un agente siga T con éxito significa que consigue maximizar su utilidad. Pero entonces no es posible que el objetivo de T sea por ello peor alcanzado, puesto que seguir con éxito T precisamente significa haber alcanzado del mejor modo posible los objetivos planteados por T. Lo que sí es cierto es que T es indirectamente auto-frustrante a nivel individual. Pero que una teoría sea indirectamente auto-frustrante no significa que falle en sus propios términos, precisamente porque obrar racionalmente no es parte del objetivo sustantivo de T. T no es una teoría auto-contradictoria.

Puede pensarse que estas paradojas son de todos modos una objeción contra T, que el simple hecho de que se den tales paradojas hace que una teoría resulte insatisfactoria. Sin embargo, esto no es así. Pensemos de nuevo en el caso 1. Si viviéramos en un mundo en el que todos fuéramos

²⁹ Parfit (86), p.5.

³⁰ Parfit (86), p.55.

transparentes, sería muy probable que nos encontraríamos a menudo en situaciones como la descrita en ese ejemplo. En ese mundo, si obráramos siempre de tal modo que intentáramos en cada momento conseguir lo mejor, el resultado sería que conseguiríamos un resultado muy malo. Si, por el contrario, cambiáramos nuestras disposiciones de tal modo que siempre estuviéramos dispuestos a cumplir nuestras palabras con independencia de que esto fuera o no lo mejor para nosotros, el resultado sería mucho mejor. Ahora bien, puesto que según T lo racional es hacer aquello que maximice nuestra utilidad, entonces T presumiblemente afirmaría que si hay una disposición tal que tenerla y obrar según ella maximiza nuestra utilidad, entonces es racional procurar adquirir esa disposición. Es decir, podríamos afirmar que la disposición a cumplir siempre la palabra dada es una disposición racional, a pesar de que los actos que se siguen de tal disposición pueden ser en ocasiones irracionales.

No hay nada especialmente paradójico en esa afirmación. Lo paradójico está en afirmar que un individuo que obra irracionalmente esta siguiendo con éxito una teoría acerca de la racionalidad. Sin embargo, T no afirma tal cosa. En el caso 1, T afirma que el individuo obra racionalmente al convertirse en una persona que cumple su palabra, aunque el cumplirla sea irracional. En el caso 2, T afirma que el individuo obra racionalmente al tomar la droga(2), aunque su comportamiento bajo los efectos de esta droga sea irracional. La acción de tomar la droga(2) o de convertirse en un individuo digno de confianza maximizan la utilidad del individuo en esas situaciones.

La objeción desaparece si recurrimos al concepto de racionalidad indirecta. Recordemos una vez más la primera condición, según la cual atarse uno mismo es realizar una determinada acción en t_1 a fin de incrementar la probabilidad de realizar otra acción en t_2 siempre que este incremento de la probabilidad de la acción en t_2 sea el motivo de la acción en t_1 . Es cierto que, tal y como Parfit afirma, T no nos pide que actuemos de un modo racional. Pero sólo en el sentido de que T no nos pide que la acción a realizar en t_2 sea una acción racional. Pero lo que sí pide T es que la acción realizada en t_1 sea racional. Esto no quiere decir que en algún sentido ser racional forme parte del objetivo sustantivo de T. Lo único que quiere decir es que el motivo de volverse irracional cuando las circunstancias lo requieren debe tener como objetivo la

consecución del objetivo sustantivo de T, e.d., la maximización de la utilidad.

Pueden aún surgir dos objeciones contra T. En *primer* lugar, puede pensarse que, puesto que seguir con éxito T significa maximizar la utilidad, entonces sólo a la vista de las consecuencias de las acciones de un individuo podremos decir si este ha seguido o no con éxito T. Pero entonces, nuestra teoría pierde todo interés, pues deja de ser una teoría normativa. Esta objeción no es válida. Es cierto que seguir con éxito T significa maximizar la utilidad, pero hay que recordar que en al definir el objetivo sustantivo de T, por "utilidad" debemos entender "utilidad esperada". Y no hace falta ver las consecuencias de las acciones de un individuo para conocer la utilidad esperada de esas acciones.

La *segunda* objeción es una variante generalizada de la anterior. Esta objeción diría: "T es una teoría que identifica racionalidad con éxito. Cuando un individuo consigue que su vida vaya lo mejor posible, entonces es un individuo que ha seguido con éxito T, y es por tanto en un sentido un individuo racional. Pero esto es ridículo. Lo realmente importante para una teoría de la racionalidad no debe ser los objetivos que se alcanzan sino *cómo* se los alcance." Puesto que esta objeción es una generalización de la anterior, la respuesta es la misma. Veamos un ejemplo similar a nuestro caso 1 y al que llamaremos caso 1'. Supongamos que no somos transparentes. Pero de repente una noche se declara una epidemia de una enfermedad nueva y desconocida a la que se bautiza como "pinochia": cada vez que mentimos nuestra nariz crece de un modo visible. Supongamos dos individuos, A y B. A es un individuo racional según T. Por el contrario, B es un individuo que, aunque bastante racional en otros aspectos, siempre cumple su palabra sin detenerse a pensar en si esto le beneficia o no. Ambos individuos han contraído la enfermedad, pero lo ignoran. Y ambos, al llegar la noche, se encuentran con que se han dejado en casa la cartera y no tienen más remedio que llamar a un taxi. Por lo tanto, se disponen a parar uno. El taxista (supongamos para simplificar que sólo hay un taxi en el sitio en el que se encuentran), al contrario que A y B, esta siempre al tanto de las últimas noticias pues, para mitigar el aburrimiento de su trabajo, siempre lleva la radio puesta. De modo que el taxista conoce la nueva enfermedad. A para el taxi. Le promete al taxista que al llegar a casa le pagará. Pero a A le empieza a crecer la nariz.

Cuanto más insiste en su promesa, más le crece. A es un individuo que nunca hace lo que es peor para él. A está mintiendo. El taxista se va. Un poco más adelante el taxista se encuentra con B. Al igual que A, B le asegura que le pagará al llegar a casa. Pero, a diferencia de A, a B no le crece la nariz. B es un hombre que siempre cumple su palabra. B está diciendo la verdad. A consigue un resultado muy malo y B consigue un resultado mejor.

¿Significa esto que podemos afirmar que A no ha seguido con éxito T mientras que B sí lo ha hecho? Ni mucho menos. Más bien todo lo contrario. A ha seguido con éxito T y B no lo ha hecho. Un agente sigue con éxito T si con su conducta maximiza su utilidad esperada. Y esto es precisamente lo que hace A. A ha tenido en cuenta la probabilidad de que el taxista se dé cuenta de que no va a pagarle y que atribuya esto a su carácter racional. Pero esta probabilidad es muy remota. A ha seguido con éxito T. Lo único que ha ocurrido es que ha tenido mala suerte. Pero la conducta racional no es un antídoto contra la mala suerte. Por el contrario, B no ha seguido con éxito T. Su comportamiento no maximizaba su utilidad esperada. Pero ha tenido buena suerte.

B es un hombre digno de confianza y, en la situación descrita, consigue el mejor resultado en términos de autointerés, e.d., el objetivo sustantivo propuesto por T. Comparemos esto con el agente del caso 1 original (p.185), al que llamaremos C. C se comporta exactamente igual que B y consigue exactamente lo mismo. Pero hay una diferencia. C se comporta como un hombre digno de confianza porque se ha transformado a sí mismo, deliberadamente, en un hombre de este tipo. Pero esto no es lo esencial. Después de todo, es posible que B también se haya transformado a sí mismo, de manera deliberada, en un hombre de confianza. Lo esencial es que C ha llevado a cabo esta transformación porque, en las determinadas circunstancias en las que se encuentra, tiene buenas razones para suponer que esto es lo que le conducirá a la obtención de los mejores resultados posibles. Es decir, C se ha transformado en un individuo que no siempre obra de manera racional precisamente porque es un individuo racional, porque realizar esa transformación era lo racionalmente recomendable. Por eso, aun haciendo lo mismo y consiguiendo lo mismo, C es un individuo racional y A no lo es.

En resumen, T puede recomendar en determinadas circunstancias ac-

tuar de un modo irracional. Esto sucede en aquellas circunstancias en las cuales actuar directamente para alcanzar el mejor resultado posible según T resulta contraproducente, e.d., en aquellas circunstancias en las cuales T es indirectamente auto-frustrante a nivel individual. En tales circunstancias un individuo racional se encuentra en una mala posición. La única salida es actuar de un modo irracional. Pero esto no es posible para un individuo racional. Por eso, este individuo hará lo mejor, en términos de T, si consigue atarse a sí mismo de tal modo que se convierta en un individuo irracional. Después de transformarse a sí mismo, será un individuo irracional que hará cosas irracionales. Pero en un sentido fundamental, es un individuo racional, a saber, en el sentido de que la decisión tomada en t_1 de hacerse irracional en t_2 es una decisión racional. En los casos en los que T es indirectamente auto-frustrante, un individuo racional seguirá la estrategia indirecta de atarse a sí mismo con el fin de que, llegado el momento, no tenga la "tentación" de comportarse de un modo racional.

CAPITULO XI

UNA TEORIA ALTERNATIVA

En el capítulo X hemos visto como nuestra teoría podía ser en ocasiones indirectamente auto-frustrante en el nivel individual. Pero, tal y como habíamos visto en capítulos anteriores, el mayor problema de nuestra teoría es que es directamente auto-frustrante en el nivel colectivo. Puesto que se trata de una teoría sobre la racionalidad individual, este fracaso a nivel colectivo no significa que nuestra teoría sea auto-contradictoria o que falle en sus propios términos. Pero, a pesar de eso, su fracaso a nivel colectivo la convierte en una teoría en algún sentido insatisfactoria. El objeto de este capítulo es analizar la posibilidad de encontrar una teoría alternativa sobre la racionalidad individual aceptable y que no presente este tipo de problemas en el nivel colectivo. Para ello, trataremos en primer lugar de localizar exactamente donde reside el problema, es decir, analizaremos nuestra teoría a fin de aislar aquella o aquellas de sus características que la hacen directamente auto-frustrante en el nivel colectivo.

I.

Como se recordará, decimos que una teoría es directamente auto-frustrante a nivel colectivo cuando ocurre que, si todos seguimos con éxito la teoría entonces, y debido a ello, sucede que los objetivos dados por la teoría a cada uno de nosotros se alcanzan peor de lo que hubiera sido posible si ninguno de nosotros hubiera seguido con éxito la teoría. Esto es lo que ocurre precisamente en el caso de los dilemas que hemos analizado en capítulos anteriores, en los que suponíamos que los agentes involucrados en la situación eran racionales según T. Supongamos una de estas situaciones representada por la siguiente matriz

	B ₁	B ₂
A ₁	(1,1)	(3,0)
A ₂	(0,3)	(2,2)

Ejemplo 11

Según T, cada agente actuará racionalmente si escoge la estrategia cuya utilidad esperada es mayor. Es decir, el agente A actuará racionalmente si escoge la estrategia A₁ y el agente B si escoge la estrategia B₁. Además, al contrario de lo que ocurría en los casos analizados en el capítulo anterior, si los agente escogen estas estrategias estarán siguiendo con éxito T, es decir, de hecho estarán maximizando su utilidad esperada individual mediante su conducta. Sin embargo, y precisamente porque los agentes siguen con éxito T, el resultado es peor para ambos de lo que hubiera sido si ninguno de ellos hubiera seguido con éxito T y hubieran escogido A₂ y B₂ respectivamente.

Estos casos pueden surgir porque T es, utilizando términos de Parfit¹, una teoría relativa al agente, e.d., porque T propone un objetivo distinto a cada uno de los agentes. Esto puede verse comparando T con una teoría que fuera neutral respecto al agente. Imaginemos que tenemos una teoría, a la que llamaremos T', que propone a todos los agentes un único objetivo, por ejemplo, que el resultado sea en conjunto lo mejor posible. En la situación representada más arriba, los agentes estarán si-

¹ Parfit (196)

guiendo con éxito T' si escogen A_2 y B_2 respectivamente. Pero si ambos agentes siguen con éxito T' , entonces estarán alcanzando del mejor modo posible el objetivo propuesto por T . Lo importante es darse cuenta de que esto sucede así necesariamente. No es posible que, si todos seguimos con éxito T' el resultado sea peor que si no lo hacemos. Seguimos T' con éxito si, entre todas las estrategias posibles, escogemos la que hace que el resultado sea el mejor posible. Y si el resultado conseguido es el mejor, entonces estamos siguiendo con éxito T' . Es decir, el par (A_2, B_2) representa al mismo tiempo y necesariamente el objetivo de T' y el resultado del seguimiento exitoso de T' . Por el contrario, en una teoría relativa al agente es posible que se den situaciones en las cuales el resultado del seguimiento de la teoría por parte de todos los agentes y el resultado más deseable según la teoría no coinciden, tal y como sucede en nuestro ejemplo, en donde el resultado del seguimiento exitoso de la teoría está representado por el par (A_1, B_1) mientras que el mejor resultado en términos de la teoría es (A_2, B_2) . Por tanto, cuando una teoría es relativa al agente, bajo ciertas condiciones la teoría resulta directamente auto-frustrante a nivel colectivo. Parfit² resume estas condiciones en dos, a saber,

Condición positiva que cada agente pueda elegir entre 1) procurarse a sí mismo el menor de dos beneficios o 2) proporcionarle al otro el mayor beneficio

Condición negativa que la elección que realice cada uno no sea en ningún otro sentido ni mejor ni peor para él³.

Quando se cumplen estas condiciones, y los agentes son T-rationales, debido al carácter relativo al agente de T , surgirán los dilemas. e.d., bajo estas condiciones T es directamente auto-frustrante a nivel colectivo. Parece entonces que sí podemos encontrar una teoría alternati-

² Parfit (86), p.57.

³ Todo esto es en realidad una manera distinta de decir cosas que ya sabíamos. Al decir que los dilemas sólo pueden surgir en una teoría que sea relativa al agente sólo estamos señalando el hecho de que los dilemas no pueden surgir en situaciones que puedan ser representadas como un juego de intereses idénticos. La condición positiva señala que los dilemas no pueden surgir tampoco en situaciones que puedan ser representadas como un juego de intereses opuestos. Es decir, los dilemas sólo pueden surgir en los juegos de intereses mixtos, en los cuales la cooperación consigue un resultado óptimo pero es individualmente inaccesible. Por otro lado, la condición negativa se limita a subrayar el hecho de que la estrategia cooperativa puede ser dominante en determinados casos de dilemas iterativos.

va aceptable que no sea relativa al agente tendremos una teoría que no fracasará en el nivel colectivo.

El análisis de T que acabamos de realizar sigue las pautas marcadas por Parfit en *Reasons and Persons*. Sin embargo, Parfit no realiza el análisis sobre T, sino sobre otra teoría acerca de la racionalidad a la que llama *Self-interest Theory* (S). Al igual que T, S es una teoría relativa al agente y que, bajo las mismas condiciones, es también directamente auto-frustrante a nivel colectivo. A pesar de esta similitud, T y S difieren en algunos aspectos relevantes. Y puesto que todo el análisis crítico de Parfit está dirigido contra S, debemos saber si T es un tipo de S o si al menos se parecen lo suficiente para que las críticas dirigidas contra S, o al menos la mayoría de ellas, puedan emplearse también contra T.

Parfit define S atendiendo a su objetivo sustancial. S propone a cada agente como objetivo de su conducta "los resultados que serían mejores para él, y que harían que su vida transcurriera para él de la mejor manera posible"⁴. Sin lugar a dudas, este es también el objetivo de T y, en este sentido, podemos decir que T es un tipo de S. Sin embargo, T no es ninguna de las variantes de S analizadas por Parfit. Según él, las distintas versiones de S corresponden a distintas posibles respuestas a la pregunta acerca de qué es o en qué consiste el interés de cada uno. Es decir, cada versión de S estaría basada en una teoría distinta acerca del interés propio. Entre ellas hay una que se parece bastante a T, a saber, aquella que afirma que el interés de cada individuo queda definido por sus preferencias personales⁵. Sin embargo, entre esta versión de S y T hay una diferencia sustancial. Esta consiste en que para T, tal y como vimos en el capítulo II, no todas las preferencias que un individuo tiene de hecho son auténticas preferencias y, más aun, no todas las preferencias auténticas de un individuo son preferencias racionales. Para T, el objetivo de la acción racional no es la maximización de la utilidad definida sobre las preferencias de hecho de un individuo, sino definida sobre el conjunto racional de sus preferencias racionales. Podemos decir que T

⁴ Parfit (86), p.3.

⁵ Esta versión de S es llamada por Parfit "Preference-Hedonism". Para un análisis de esta versión ver Parfit (86), Apéndice 1.

es una versión crítica de S, en el sentido de que T discrimina críticamente entre las preferencias de un individuo, eliminando algunas de ellas como irracionales y considerando que otras son racionalmente exigibles*.

Como hemos dicho en repetidas ocasiones, el hecho de que T sea directamente auto-frustrante a nivel colectivo, a pesar de hacerla insatisfactoria en algún sentido, no es una objeción contra ella, puesto que es una teoría acerca de la racionalidad individual que no pretende resultar satisfactoria a nivel colectivo. Sin embargo, Parfit cree que S (y T en tanto que es una versión de S) puede ser criticada desde una teoría alternativa acerca de la racionalidad individual. La segunda parte de este capítulo estará dedicada al análisis de esta teoría alternativa, sus diferencias con T y el desafío que esta alternativa presenta a T.

* Parfit no menciona en absoluto la posibilidad de versiones críticas de S del tipo de T. Sin embargo, cuando presenta su teoría alternativa acerca de la racionalidad (P) sí presenta versiones críticas. De hecho, la teoría defendida por él es una versión crítica de P. No conviene olvidar sin embargo que son posibles versiones críticas de S y que nuestra teoría es una de ellas, sobre todo porque algunas de los argumentos de Parfit a favor de P y en contra de S son simplemente argumentos a favor de una teoría crítica y por lo tanto no descalifican a T.

II. La teoría de los objetivos presentes

Parfit presenta *S* caracterizándola por el objetivo que esta teoría propone y sólo después presenta las distintas versiones que pueden darse de esta teoría. Sin embargo, al presentar la teoría alternativa *P*, Parfit procede de un modo distinto, ofreciendo directamente las diversas versiones alternativas de *P*. Nosotros seguiremos aquí también este procedimiento y sólo después analizaremos que es lo que todas estas versiones tienen en común. Las tres distintas versiones de *P* aparecen caracterizadas por su objetivo sustantivo, e.d., por lo que cada una de ellas afirma que constituye una razón para actuar:

Teoría instrumental (IP), según la cual lo que constituye para cada uno una razón para actuar es aquello que satisface en el mayor grado posible sus deseos presentes.

Teoría deliberativa (DP), según la cual lo que da a cada uno una razón para actuar es lo que alcanza del mejor modo posible, no lo que uno desea realmente, sino lo que uno desearía en el momento de actuar si hubiera pasado por un proceso de deliberación ideal, es decir, si conociera todos los hechos relevantes, estuviera pensando claramente y no estuviera sometido a ningún tipo de influencia distorsionante.

Teoría crítica (CP), que afirma que 1) algunos deseos son intrínsecamente irracionales, 2) un conjunto de deseos puede ser irracional incluso si los elementos del conjunto no lo son, debido a las relaciones formales que estos elementos mantienen entre sí, 3) un conjunto de deseos puede también ser irracional porque no contenga ciertos deseos que están racionalmente exigidos. Si suponemos que mis deseos han sobrevivido a un proceso de deliberación ideal y que el conjunto de mis deseos no es irracional, entonces lo que me ofrece una razón para actuar es aquello que mejor satisface aquellos de mis deseos presentes que no son irracionales. Si recordamos nuestra presentación de *T* en los capítulos I y II veremos que estas tres versiones de *P* se parecen extraordinariamente a los tres pasos que dimos para definir la acción racional según *T*. En un primer paso dijimos que era aquella que maximizaba la utilidad definida sobre el conjunto de preferencias del agente, lo cual corresponde a la afirmación de IP. En segundo lugar, distinguimos entre las preferencias de hecho de un agente y sus preferencias auténticas, identificando estas últimas con

aquellas que el agente tendría si hubiera realizado un proceso de deliberación ideal, afirmación que se corresponde a la mantenida por DP. Por último, vimos que era necesario 1) exigir que las distintas preferencias de un agente cumplieran determinados requisitos formales, 2) rechazar como intrínsecamente irracionales determinados tipos de preferencias, 3) considerar que un conjunto de preferencias, si contiene determinados elementos, debe entonces considerarse como un requisito de racionalidad la presencia de algunos otros. Estas exigencias corresponden a las realizadas por CP. Nuestra definición final de lo que podía considerarse acción racional admite por tanto todas las exigencias presentadas por CP (las exigencias de DP, que también son exigencias de T, están recogidas en CP). Entonces, ¿qué es lo novedoso en P?, ¿qué es lo que distingue P de T?

Parfit presenta las versiones de P como distintas a las versiones de S fundamentalmente porque no considera ninguna versión crítica de S, sino únicamente versiones instrumentales que corresponderían a la versión IP de P. Esto le lleva a presentar DP y CP como absolutamente distintas. Sin embargo, T es una versión crítica de S, que se correspondería con CP. Debemos tener esto en cuenta a la hora de analizar las diferencias entre nuestra versión de S, (T), y P. Pero entonces, si la diferencia entre T y P no es que esta última teoría ejerza una crítica racional que discrimine entre las distintas preferencias que de hecho tiene un agente, pues esto es también una característica de T, ¿cuál es la diferencia?

Si nos fijamos en las distintas versiones de P, veremos que en todas ellas se hace referencia a los deseos *presentes* del agente. Esto parece oponerse a la exigencia de neutralidad temporal realizada por T. Una posibilidad es que la diferencia entre ambas teorías resida en este punto, tal y como defienden algunos autores, por ejemplo, Kagan⁷. La teoría de la racionalidad práctica tiene que ver con los deseos y preferencias que nos dan una razón para actuar. Y lo que distingue unas de otras es el tipo de deseos que según cada una nos ofrecen estas razones

"Imaginemos que me presentan una matriz completa de información, en la cual aparecen listados todos los deseos de las personas pa-

⁷ Kagan (1986), pp. 745-50.

sados presentes y futuros. Cada deseo aparece indexado de tal modo que se indica de quién es el deseo (si es mío, tuyo o de Fritz) y cuando ocurre (si ahora, ayer por la tarde o dentro de veinte años). Dada esta abundancia de información, ¿cuál es para mí la acción racional a realizar? ¿a qué deseo debo dar peso en mis deliberaciones? ¿Y cuanto peso?"".

Las distintas respuestas a estas preguntas dan lugar a distintas teorías acerca de la racionalidad práctica. La respuesta de cada teoría se concreta en lo que Kagan llama la "Función Maestra" de cada una de ellas. Concretamente, en las respuestas a estas preguntas está la diferencia entre S y P. S afirma que debo dar peso únicamente a aquellas preferencias que son *mías*. Sólo estas nos ofrecen una razón para actuar. Nuestra versión crítica de S (T) afirmaría que debemos dar peso únicamente a aquellas preferencias *mías* que son racionales. Por otra parte, P afirmaría que debemos dar peso exclusivamente a aquellas preferencias que son *mías ahora*. La versión crítica de P, CP, diría que debemos dar peso a aquellas preferencias *mías ahora* que son racionales. Por consiguiente, para Kagan la diferencia entre S y P está exclusivamente en la aceptación o rechazo de las consideraciones temporales.

Sin embargo, Parfit rechaza expresamente esta afirmación⁹. Según él, la diferencia entre S y P no es el rechazo o la admisión de la neutralidad temporal. Es posible una versión de CP que admita la neutralidad temporal como un requisito de racionalidad. Esta versión haría que CP coincidiera parcialmente con S, pero aun así, para Parfit ambas teorías seguirían siendo distintas. Supongamos que CP afirma lo siguiente:

CP1 "cada uno de nosotros está racionalmente requerido a preocuparse de su propio interés. Y esta preocupación debe ser temporalmente neutral. Cada uno de nosotros debe estar igualmente preocupado por todas las partes de su vida. Pero, a pesar de que todos debemos tener esta preocupación, no es necesario que esta sea la dominante"¹⁰.

Ahora bien, si tenemos una versión de CP que afirma esto, ¿hasta que punto esta versión no es equivalente a T? Parfit es consciente de que existe el peligro de pensar que S coincide con P. De hecho, ambas teorías

⁹ Kagan (196), p. 746.

¹⁰ La sección 32 de *Reasons and Persons* está totalmente dedicada a argumentar contra esta afirmación.

¹¹ Parfit (196), p. 135.

coinciden bajo el supuesto del egoísmo psicológico. Este afirma que lo que una persona desea, supuesto un proceso de deliberación ideal, es aquello que es mejor para él, o aquello que promueve mejor su interés a largo plazo. De modo que el mejor modo de analizar si hay diferencias entre S y P es considerar si puede aceptarse ese supuesto. El egoísmo psicológico puede hacerse verdadero por definición de tres modos distintos; puede afirmarse que

1- si alguien desea lo que es peor para él, entonces no puede estar pensando claramente, y debe estar sometido a alguna influencia distorsionante;

2- lo que es mejor para alguien es por definición lo que él desea, o desearía, si hubiera realizado un proceso de deliberación ideal;

3- si un acto satisface los deseos presentes de un agente entonces, por definición, ese acto maximiza la utilidad de ese agente.

La primera de estas afirmaciones hace que P pierda su independencia y coincida con S de forma trivial. La segunda hace coincidir a S con DP. Una versión de S que admitiera esta afirmación estaría definiendo el interés propio en términos de lo que cada uno desea, supuesta la deliberación ideal. Por último, la tercera haría coincidir a S por definición con IP. Puesto que nuestra versión de S es una versión crítica, parece que no podemos defender ninguna de estas afirmaciones, pues T sólo puede coincidir con CP, pero en ningún caso con las otras versiones de P. Sin embargo, podríamos hacer la afirmación del egoísmo psicológico verdadera por definición de otra forma, por ejemplo, podríamos sostener

4- lo que es mejor para alguien es, por definición, lo que satisface, no sus deseos de hecho, sino aquellos de sus deseos que son auténticos y racionales.

Esta afirmación haría que S coincidiera con CP. Es decir, haría T y CP equivalentes por definición. Debemos por tanto considerar si podemos mantener esta afirmación. Si no podemos mantenerla, parece que entonces hay una diferencia real entre T y CP que debemos localizar. Desde luego, la afirmación (4) siempre puede mantenerse si la tomamos como una definición de lo que entendemos por "ser mejor para alguien". Pero, entendida de esta forma, (4) deja de ser una afirmación interesante. Es decir, si es esta la forma en que definimos en T "ser mejor para alguien", entonces simple y trivialmente, T es CP en el sentido de que nuestra teo-

ría es realmente CP formulada en un modo confuso, ya que habitualmente no entendemos que "ser mejor para alguien" signifique esto. En este caso haríamos mejor en reconocer que no hay diferencia entre ellas y definir nuestra teoría como CP. Mantener que hay una diferencia entre T y CP equivale a mantener que la afirmación (4) sólo es verdadera si se entiende como una definición, es decir, mantener que en el sentido habitual de "ser mejor para alguien" puede suceder que las preferencias racionales de la gente no siempre estén ligadas a lo que es mejor para ellos. Y defender T frente a CP equivaldría a afirmar que, cuando las preferencias racionales de un agente no coinciden con lo que es mejor para él, entonces lo racional es actuar según lo que es mejor para él. Según T, una acción racional es aquella que maximiza la utilidad del agente definida sobre el conjunto de sus preferencias racionales. Esto es, por otra parte, lo que afirma (CP1). Si esto es todo lo que afirma T, entonces T no es una versión de S a menos que cometamos el error (o mejor, la confusión) de aceptar (4) como una definición.

Podemos empezar por preguntarnos si realmente hemos definido "ser mejor para alguien" del modo indicado en (4) y si, por ello, nuestra teoría es trivialmente equivalente a CP y no es en absoluto una versión de S. Para T los intereses de alguien están expresados en el conjunto de sus preferencias racionales. Por tanto, hay un sentido en el que lo mejor para alguien equivale a la maximización de la utilidad definida sobre el conjunto de sus preferencias racionales, a saber, en el sentido de que lo mejor para alguien es que se satisfagan sus intereses. Por supuesto esto es una definición, pero no es una definición de la teoría sino una parte de lo que realmente entendemos por ser mejor para alguien. El problema está en que hay otro sentido de "ser mejor para alguien" según el cual esta expresión hace referencia a los intereses egoístas del agente. Como ya sabemos, T no supone que los intereses de alguien son necesariamente egoístas. Cuando S afirma que los intereses de los demás no tienen ningún peso en la deliberación de un agente racional debe entenderse que nos estamos refiriendo a "peso directo". Kagan expresa esto de un modo extremadamente claro

"La teoría del interés propio (S) da pleno peso a cada uno de mis propios intereses (...) y ningún peso directo a los intereses de los demás. Digo "peso directo" porque si uno de mis deseos es, por ejemplo, ver que mi madre es feliz, entonces sus deseos entrarán

en mis deliberaciones indirectamente; pero no tendrán ningún peso por derecho propio. Según S, si no fuera por mi interés en ella, los deseos de mi madre no entrarían en mis deliberaciones. Sin embargo, cada uno de mis propios deseos genera directamente una razón para actuar¹¹.

Sin embargo, esta posible confusión, que ya hemos aclarado repetidamente, no puede ser la responsable de la diferencia entre S y CP, puesto que Parfit reconoce que para S el interés propio no está necesariamente definido en términos egoístas¹². Para T, el interés propio está definido en términos de lo que hemos llamado "preferencias personales" y, en este sentido y sólo en este, T afirma que lo mejor para alguien es ver satisfecha del mejor modo posible sus preferencias personales. T es por tanto una versión crítica de S. Ahora bien, esta definición de T de lo que es mejor para alguien es distinta de la definición (4). Mientras que la identificación de T de lo que es mejor para alguien es admisible, la realizada por (4) es, o bien una definición artificial y, por ello trivial, o bien falsa. Cuando se entiende (4) no como una definición sino como una afirmación substantiva, resulta falsa porque

"La mayoría de nosotros, en la mayoría de las ocasiones, deseamos actuar en nuestro propio interés. Pero hay muchos casos en los que esto no es el deseo más fuerte o en los que, incluso si lo es, está compensado por muchos otros deseos. Hay muchos casos en los que esto sucede incluso cuando el agente conoce todos los hechos relevantes y está pensando claramente. Esto sucede, por ejemplo, cuando la teoría de los objetivos presentes apoya la moralidad en conflicto con el autointerés. Lo que más desea alguien pueda ser cumplir su deber, incluso sabiendo que esto va contra sus propios intereses".¹³

En este texto, Parfit defiende la falsedad de (2). Al no considerar versiones críticas de S, no se plantea la necesidad de argumentar contra una afirmación como la de (4)¹⁴. Sin embargo, su argumento puede

¹¹ Kagan (86), p.747.

¹² Parfit (86), p.5.

¹³ Parfit (86), p.128.

¹⁴ A pesar de no considerar versiones críticas de S, en este texto se hace una referencia implícita a lo que podríamos llamar una versión semi-crítica de S, o, más claramente, a una versión deliberativa de S, correspondiente a la versión DP de P. Esta versión deliberativa de S afirmaría que lo que da a un agente una razón para actuar es aquello que satisface aquellas de sus preferencias que sobrevivirían a un proceso de deliberación ideal.

extenderse fácilmente para cubrir nuestro caso. Para argumentar la falsedad de (4) como afirmación substantiva, basta con defender que en ocasiones, incluso cuando un agente cumple todas las condiciones de racionalidad exigidas por T, es cierto que nuestro mayor deseo no es la satisfacción de nuestras preferencias personales. Es decir, lo que hace de CP una teoría auténticamente diferente de T, y lo que, sí es cierto, puede constituir un argumento a favor de CP y en contra de T, es la existencia de lo que hemos llamado "preferencias morales" junto a las preferencias personales.

Nosotros hemos reconocido la existencia de este tipo de preferencias. Su satisfacción no es mejor para el agente, a menos que hagamos coincidir trivialmente ambos conceptos por definición, pues la satisfacción de las preferencias morales de un agente no supone la satisfacción de sus intereses, definidos sobre sus preferencias personales. Por lo tanto, T no es equivalente a CP, sino que ambas son teorías realmente distintas. La diferencia entre ambas puede expresarse del modo siguiente: de acuerdo con T, la acción racional es aquella que maximiza la utilidad del agente definida sobre el conjunto de sus preferencias personales racionales, mientras que de acuerdo con CP, la acción racional es aquella que satisface del mejor modo posible las preferencias racionales del agente, incluyendo no sólo las preferencias auto-interesadas, e.d., las preferencias personales, sino cualquier preferencia racional. Una vez que hemos establecido de un modo suficientemente claro la diferencia entre T y CP, podemos pasar a considerar los argumentos ofrecidos por Parfit contra T y a favor de CP.

III. Los argumentos contra T

En cada uno de los tres argumentos que ofrece, Parfit señala una diferencia específica entre ambas teorías. La defensa de P (más concretamente, la defensa de CP), consiste en señalar que, siempre que ambas teorías difieren, CP es más defendible. Aunque Parfit siempre hace referencia a versiones no críticas de S, aquí veremos sus argumentos adaptados, siempre que sea posible, a nuestra versión crítica de S, T.

El primer argumento de Parfit

Según T, la conducta de un agente racional debe estar dirigida a la maximización de la utilidad definida sobre el conjunto de sus preferencias racionales. Es decir, un agente T-racional dará peso en sus consideraciones sólo a sus intereses, sea cual sea el posible costo que esto suponga para los intereses de los demás. Por tanto, cuando el interés propio entra en conflicto con la moralidad, T nos aconseja elegir lo primero. Siguiendo la terminología de Parfit, llamaremos a esta característica de T (y de todas las versiones de S en general) *sesgo a favor de uno mismo*.

La mayoría de nosotros tenemos habitualmente tal sesgo. Es decir, nuestras preferencias racionales son por lo general preferencias personales¹⁵. En los casos en los que esto sucede, CP coincide con T. Según ambas teorías, lo que un agente racional debe hacer en estos casos es tratar de satisfacer del mejor modo posible sus preferencias personales. Sin embargo, puede suceder que alguien, o cualquiera de nosotros en algunos momentos, no quiera dar el mayor peso, o un peso exclusivo, a sus preferencias personales. Supongamos que yo he cometido alguna falta, por ejemplo, que he escrito en el vestuario de mi estudio de baile algún comentario jocosos de dudoso gusto sobre el pianista. No hay ninguna prueba de que yo sea la culpable, de modo que si no confieso no podré ser acusada. Supongamos que la situación es la siguiente. Si confieso, seré castigada a no tomar la clase durante tres días. Si no confieso, todos (yo inclui-

¹⁵ Hay un sentido en el que todos tenemos preferencias morales, a saber, en el sentido de que si razonáramos desde el punto de vista moral, tendríamos estas preferencias. Pero la mayoría de nosotros casi nunca tiene estas preferencias en el sentido relevante, es decir, en el sentido de que queramos obrar según estas preferencias.

da, naturalmente) nos quedaremos ese día sin clase. La cosa es especialmente grave, porque faltan tres días para los exámenes del conservatorio y en esas circunstancias perder dos días de clase puede ser fatal.

Desde luego, para mí es peor estar tres días sin clase que perderla por un día. Además, no me gustaría en absoluto enfrentarme frontalmente con el pianista, y menos aún considerando que en el examen puede jugar una mala pasada, ni muchísimo menos pasar por el trance de tener que disculparme ante él. Por supuesto, siento mucho que mis compañeros sufran un castigo injustificado. Son mis amigos. Pero aún así, es peor para mí perder tres días de clase. Mi cariño por ellos no llega a tanto. Sin embargo, y a pesar de que es mejor para mí no confesar, creo sinceramente que debo hacerlo. Es más, precisamente porque creo que debo hacerlo, quiero hacerlo. En algún sentido, prefiero confesar. No es que eso tranquilice mi conciencia y aumente así la utilidad de mi acción. Por supuesto que la tranquiliza y por supuesto que la aumenta. Pero no lo suficiente como para hacer que sea mejor para mí confesar. Tampoco es que yo tenga en gran aprecio mi concepto de mí misma como una persona honrada. Lo tengo, pero no lo suficiente. Perder tres días de clase tan cerca de los exámenes es peor para mí que pensar que, después de todo, no soy tan valiente como suponía. Se mire por donde se mire, es peor para mí confesar. Pero sé que tengo que hacerlo y, por eso, prefiero hacerlo. No es una preferencia personal. Es una preferencia moral y yo quiero razonar desde un punto de vista moral.

Según Parfit, en estos casos CP y T difieren. De acuerdo con T, lo que tenemos que hacer en esos casos es actuar de acuerdo con nuestras preferencias personales, e.d., hacer lo que es mejor para nosotros. Sin embargo, según CP lo que debemos hacer depende de la intensidad relativa de nuestros intereses personales y nuestras preferencias morales. Para CP, estas últimas, supuesto que son racionales, deben tener también un peso en nuestras deliberaciones. En caso de conflicto, la decisión depende de su fuerza relativa. Y es posible que, una vez que todo haya sido sopesado, la preferencia moral sea la prevaleciente. En este caso, lo racional sería actuar según mi preferencia moral. Es decir, sería racional hacer lo que es peor para mí. Al menos, esto es así en algunas versiones de CP, aunque no en todas. Concretamente, no sería cierto en una versión de CP que afirmara que los deseos ligados al interés propio son

los deseos supremamente racionales y que es irracional dar el mismo peso a cualquier otra cosa.

Esta versión de CP, a la que llamaremos CPS, coincidiría con T en decir que, en estos casos, al igual que en todos los demás, lo racional es obrar según las preferencias personales. Por ello, podemos ver por qué estos casos en los que CP y T difieren hablan a favor de CP si analizamos por qué CPS no es una versión admisible. CPS puede presentarse en dos versiones. La versión fuerte afirmaría que las únicas preferencias racionales son las preferencias personales. Pero es posible una versión débil de CPS que admita que las preferencias morales pueden ser racionales. Según esta versión, a pesar de que las preferencias morales pueden ser racionales, no son supremamente racionales. Por ello, es racional darles algún peso en nuestras deliberaciones, pero es irracional darles el mismo peso que a las preferencias personales.

Consideremos en primer lugar si la versión fuerte de CPS es aceptable. Su aceptación depende de si realmente podemos considerar que las preferencias morales no son racionales. Supongamos que las preferencias morales de un agente son auténticas preferencias, e.d., que sobreviven a un proceso de deliberación ideal. ¿Que motivos podemos alegar para defender que no pueden ser racionales? En el capítulo II analizamos los diversos motivos por los que podía considerarse que una preferencia o un conjunto de preferencias es irracional. Empecemos entonces por analizar si un conjunto de preferencias que contenga preferencias morales puede ser racional.

Un conjunto de preferencias es racional si se mantiene entre sus miembros una serie de relaciones. En concreto, un conjunto racional de preferencias ha de cumplir las condiciones de reflexividad, transitividad y completud, e. d., ha de ser una ordenación. La reflexividad no es problemática, puesto que no hay nada que impida que las preferencias morales sean reflexivas. Las otras dos condiciones son más problemáticas, al menos aparentemente. Esto sucede porque, desde luego, no podemos decir que un conjunto de preferencias que contenga preferencias personales y preferencias morales sea transitivo o completo.

En efecto, la ordenación de un conjunto de elementos supone la presencia de una cualidad común a todos ellos que se toma como base de la ordenación. Pero si esta cualidad es la de "ser preferido personalmen-

te", entonces, por definición no es la de "ser preferido moralmente", supuesto que ambos tipos de preferencias son distintas. Supongamos el conjunto de alternativas de nuestro ejemplo y llamemos A a la alternativa "confesar" y B a "no confesar", a las cuales añadimos otra alternativa C "cambiarme de estudio de baile". Estas alternativas se relacionan entre sí de distintos modos según el atributo que se utilice para ordenarlas. Si este es "ser preferida personalmente", entonces $\{aRa, bRb, cRc, bRa, bRc, aRc\}$. Este conjunto de alternativas, relacionadas entre sí de este modo, es una ordenación. Si el atributo es "ser moralmente preferida", entonces $\{aRa, bRb, cRc, aRb, aRc, bRc, cRb\}$, lo cual también constituye una ordenación.

Esto significa que un conjunto de preferencias morales, e.d., de alternativas vinculadas según la relación "ser moralmente preferido a" puede ser una ordenación. Pero no puede hacerse una ordenación según dos relaciones distintas. Por lo tanto, si consideramos que la relación relevante es la preferencia personal, la preferencia moral no tiene cabida y resulta "irracional" un conjunto que pretenda ordenarse según ambas relaciones. Pero lo mismo sucede si se considera que la relación relevante es la de preferencia moral. Que la relación relevante sea la de ser preferido personalmente es el supuesto de T y de todas las versiones de S. Puede que ese supuesto sea defendible, pero lo que no puede hacerse en ningún caso es defenderlo alegando que, puesto que estamos hablando de preferencias personales, introducir las preferencias morales haría imposible la ordenación.

Lo que sí puede hacerse es ordenar el conjunto de alternativas según una relación que incluya ambos tipos de preferencias. Por ejemplo, podemos ordenarlas según la relación "ser más deseado que", que podemos entender como "ser lo que el agente elegiría una vez consideradas todas sus preferencias personales y morales". La cuestión relevante es entonces si un conjunto ordenado según esta relación puede ser una ordenación. Y la respuesta es sin duda afirmativa.

Podemos pasar ahora a considerar si las preferencias morales aisladas pueden ser racionales. Tal y como vimos, una preferencia puede ser intrínsecamente irracional si consiste en preferir lo peor. De nuevo nos enfrentamos a un problema parecido al anterior. Si lo racional es obrar según el interés propio, entonces elegir según las preferencias morales

en los casos en los que ambas cosas están en conflicto, es irracional. Desde el punto de vista del interés propio, elegir moralmente es elegir lo peor. Pero desde el punto de vista de la moral, elegir moralmente es lo mejor. Lo que es irracional en general es elegir lo que yo considero peor en general, pero no necesariamente lo que es peor para mí, entendido esto como "peor para mí desde el punto de vista de mis intereses personales". Puede ser que yo considere que es mejor elegir moralmente en determinados casos. Si es así, ¿qué motivos hay para mantener que esta preferencia es irracional? No es preferir lo peor. Mas bien, es preferir lo mejor.

Esto puede expresarse, tal y como lo hicimos en el capítulo I, en un lenguaje neutro que no presuponga la aceptación de T o de CP diciendo que es irracional en general preferir lo peor sin ninguna razón. Preferir perder dos días de clase a preferir perder un día las diagonales es preferir lo peor. Por tanto, esta preferencia es irracional si no puedo dar ninguna razón. Pero no lo es si puedo decir que creo que es mejor vivir aceptando las consecuencias de los propios actos. Esta explicación puede entenderse en términos de preferencias personales, si considero que no aceptar las consecuencias de mis acciones disminuye el valor de mi vida mas de lo que lo hace el perder dos días de clase. Pero también puede ser una explicación moral si considero que es preferible satisfacer un determinado criterio moral a fomentar mis intereses personales.

En cualquiera de los dos casos, mi preferencia no es una preferencia de lo peor sin ninguna razón. Hay una razón. Hacer que las preferencias morales no sean racionales supone mostrar que las razones morales no son buenas razones. Pero naturalmente no puede argumentarse que no son buenas razones desde el punto de vista del interés personal. Ni lo son ni pretenden serlo. Las razones que apelan al interés personal tampoco son buenas desde el punto de vista moral. No hay por tanto ningún motivo teóricamente neutral para desechar las preferencias morales como irracionales.

Podemos entonces rechazar la versión fuerte de CPS. Aun así, podemos argumentar tal como lo hace la versión débil de CPS, diciendo que, aunque no sea irracional dar peso a las consideraciones morales, sí lo es darles el mismo peso o más que a las consideraciones personales. Esta versión débil tiene la ventaja de no cometer el error teórico de preten-



der que las preferencias morales son irracionales, pero a nivel práctico es exactamente equivalente a la versión fuerte, es decir, ambas afirman que, en caso de conflicto, un agente racional debe actuar según sus preferencias personales.

La versión débil admitiría que un agente racional puede en ocasiones dar peso a consideraciones morales, pero sólo cuando él es personalmente indiferente entre las distintas alternativas, o bien cuando la acción moralmente recomendada es también la que mejor satisface sus intereses. Pero al afirmar que ocuparse de los intereses personales es supremamente racional y que no es racional dar tanto peso a ninguna otra cosa, la versión débil afirma que en caso de conflicto no sería racional dar a ambas partes el mismo peso. Es decir, nunca sería racional obrar en contra del interés propio.

Esto ocurriría incluso en casos extremos. Supongamos que tenemos dos alternativas X e Y, que representan dos situaciones alternativas para dos individuos A y B. Las utilidades de estos individuos en estas situaciones son $X=(4,0)$ e $Y=(3,9,3,9)$. Supongamos que yo soy el individuo A. Entonces, según la versión débil de CPS, yo obraría irracionalmente si escogiera la alternativa que mejor satisface mis preferencias morales, a saber, la alternativa Y, pues esto equivaldría a dar más o igual peso a estas que a las consideraciones auto-interesadas. Y esto sucede a pesar de que las preferencias personales no son irracionales y de que yo tengo de hecho esas preferencias. Es más, incluso si mi mayor deseo es satisfacer mis preferencias morales, según la versión débil de CPS sería irracional hacerlo.

Es posible plantear variantes de la versión débil de CPS en las cuales no puedan darse estos casos extremos. Una de las posibilidades, a la que llamaremos CPS1, es entender la afirmación fundamental de CPS como diciendo que pueden dársele a las consideraciones morales un peso determinado, inferior al que se da a las preferencias personales, y que es irracional darles más. Por ejemplo, podría afirmar que a las consideraciones no personales no debe dárseles un peso superior a 1/10 del peso que se da a las consideraciones personales. Pero una teoría de este tipo, aparte de compartir las objeciones que pueden hacérsele a nuestra versión débil de CPS, tendría que justificar de algún modo esta proporción, y no parece que sea fácil hacer esto.

También podría presentarse una variante de CPS que dejará en manos del agente cuánto peso debe darse a las consideraciones morales frente a las consideraciones personales, exigiendo únicamente que este peso sea siempre menor. Esta versión de CPS, a la que nos referiremos como CPS2, sólo exigiría que, en caso de igualdad, se decidiera a favor de la alternativa mas favorable desde el punto de vista personal. De este modo, en el caso de nuestro ejemplo, esta versión permitiría que yo confesara (aunque también admitiría que no lo hiciera si es que decido dar mucho menos peso a las consideraciones personales), al menos si el número de mis compañeros es lo suficientemente grande como para que la diferencia de utilidad media entre confesar y no confesar sea considerablemente mayor que la diferencia de mi utilidad personal entre ambas alternativas. En un caso más claro, en la alternativa de romperme una uña y salvar la vida a un niño, esta versión permitiría que escogiera lo segundo sin tacharme de irracional. Sin embargo, en caso de que mis compañeros de baile fueran sólo los suficientes como para hacer que lo que pierdo entre todos si no confieso no sea mayor que lo que pierdo yo sola si confieso, esta versión afirmaría que es irracional por mi parte confesar¹⁴.

Esta variante tiene sobre la anterior la ventaja de no tener que justificar un porcentaje arbitrario. Y tanto CPS1 como CPS2 se diferencian de la primitiva versión débil de CPS en que no coinciden siempre en sus implicaciones prácticas con la versión fuerte. Sin embargo, todas las versiones débiles de CPS están sujetas a la misma objeción fundamental. Todas las versiones de CPS y, por tanto, todas las versiones de S, son injustificadamente restrictivas e intolerantes. No hay ningún motivo para suponer que las preferencias morales tengan que ser necesariamente irracionales, ni tampoco para suponer que el deseo de actuar desde un punto de vista moral sea irracional.

¹⁴ Puede pensarse que esta versión de CPS coincidiría con nuestra teoría acerca de la moralidad, pues en esta también sería preferible no confesar en este caso, pues ello produce la partición más igualitaria posible sin pérdida de utilidad. Dejando aparte la cuestión de si este veredicto es o no objetable, es importante notar que ambas teorías no son por ello iguales. En CPS, no confesar es racional porque es la más favorable a mi interés, mientras que desde el punto de vista moral que no confesar sea o no mejor que hacerlo no depende en absoluto de mi interés y ni siquiera de quién sea yo. Además, ambas teorías tampoco coinciden siempre en sus implicaciones prácticas, a menos que CPS afirme que no está en manos del agente decidir cuándo debe obrar según consideraciones morales, sino que debe hacerlo siempre y cuando su pérdida de utilidad personal no sea mayor a la pérdida de utilidad media.

Pero, si admitimos esto, tal y como lo hacen las versiones débiles de CPS y como lo hace T, en tanto que las preferencias morales y el deseo de actuar moralmente pueden cumplir los requisitos de racionalidad exigidos por T, entonces ¿qué motivos podrían aducirse para defender que es irracional actuar según estas preferencias? La afirmación de CPS de que los deseos auto-interesados son supremamente racionales es arbitraria. CPS no ofrece ninguna razón para defender tal afirmación. Mas aun, difícilmente podría ofrecerla, pues no parece que pueda haber ningún motivo para justificar que es irracional actuar según unas preferencias racionales, ni siquiera para defender que actuar según unas preferencias racionales es más racional que actuar según otras. Por tanto, el primer argumento de Parfit contra T afirma que no es defendible mantener que el sesgo a favor de uno mismo es supremamente racional, y que defender esta postura es arbitrariamente restrictivo. Tal sesgo no es supremamente racional. Por el contrario, debemos aceptar la versión de CP que afirma

(CP2) Hay al menos un deseo que no es irracional, y que no es menos racional que el sesgo a favor de uno mismo. Este deseo es el de actuar según los intereses de otras personas cuando actuar así es o bien moralmente admirable o bien un deber moral.¹⁷

Una versión de CP que realizara esta afirmación no cometería el error de escoger arbitrariamente uno de los posibles deseos racionales y elevarlo a un rango superior, es decir, esta versión de CP sería tolerante donde S es arbitrariamente intolerante. Esto constituye una razón a favor de P y en contra de S.

Algunos autores han encontrado este primer argumento de Parfit objetable. Por ejemplo, Kagan¹⁸ argumenta que en un sentido S es en efecto intolerante, pero en un sentido en el que P también lo es. Es más, en este sentido toda teoría acerca de la racionalidad es intolerante. Este sentido no parece por tanto relevante. Y en otro sentido, ni S ni P son intolerantes.

Veamos en primer lugar el sentido en el que ambas teorías son intolerantes. Toda teoría de la racionalidad se caracteriza por lo que Ka-

¹⁷ Parfit (1963), p.131.

¹⁸ Kagan (1966), pp.250-2.

gan llama su función maestra¹⁹. Y toda teoría de la racionalidad es intolerante en el sentido de que para cada una actuar racionalmente es actuar en conformidad con su función maestra. Es decir, para S el deseo supremamente racional para un agente es el de dar peso únicamente a sus intereses, del mismo modo que para P lo es el de dar peso exclusivamente a sus intereses ahora. Llamemos a este deseo de actuar en conformidad con la función maestra "metadeseo". No ser intolerante respecto al metadeseo equivaldría a no defender ninguna función maestra en concreto, lo cual sería tanto como no defender ninguna teoría de la racionalidad en absoluto.

Puede decirse por tanto que S es en este sentido intolerante. Pero también lo es P y lo es cualquier otra teoría posible. Difícilmente por tanto puede constituir esto una crítica. Y no sólo porque esta crítica podría hacerse extensiva a cualquier teoría, sino porque no es de recibo criticar a una teoría bajo el cargo de ser intolerante al defender esa teoría y no cualquier otra. Por ello, este sentido de "intolerancia" no sólo es universal sino también irrelevante.

Hay otro sentido relevante de intolerancia que sí podría constituir una crítica para una teoría que lo fuera. Es el que hace referencia, no al metadeseo, sino a los deseos de primer orden admitidos por la teoría. Pero en este sentido ambas teorías son tolerantes. Ninguna de ellas rechaza arbitrariamente ningún tipo de deseo ni afirma que algunos de ellos sean supremamente racionales. S da peso a todos los deseos del agente y P a todos los deseos actuales del agente, sin hacer entre ellos ninguna discriminación arbitraria²⁰. La crítica de Kagan se resume entonces en que 1) en un sentido irrelevante, ambas teorías son intolerantes y 2) en un sentido relevante, ninguna lo es.

Kagan tiene razón al afirmar que, en lo que se refiere a los de-

¹⁹ Ver p.222 mas arriba.

²⁰ Esta afirmación necesita ser matizada. Las versiones críticas de S y P, T y CP, sí rechazan algunos deseos, a saber, los deseos o conjuntos de deseos que no satisfacen alguno de los requisitos de racionalidad impuestos por estas teorías. Pero en tanto que sólo rechazan estos, ambas teorías serían tolerantes (si Kagan tiene razón) en el sentido que ninguna de ellas rechazaría, ni elevaría a un nivel distinto de los demás, ningún deseo que se admita como racional. La única excepción sería la versión de S que se apoya en la teoría del auto-interés que hemos llamada "Lista Objetiva" y que sí escoge determinadas cosas como los únicos objetivos racionales.

seos de primer orden, ninguna de las teorías es intolerante. Una teoría intolerante en este sentido sería, por ejemplo, la versión fuerte de CPS. Ni T ni CP son intolerantes en este sentido. Ambas admiten que las preferencias morales pueden ser racionales. También tiene razón en señalar que, con respecto a la función maestra, toda teoría es necesariamente restrictiva. Pero la crítica de Parfit no hace referencia a esto, sino al hecho de que la función maestra de T es injustificadamente restrictiva, dado que, aún siendo tolerante con respecto a los deseos de primer orden, no admite como racional la actuación según algunos de estos deseos. Parfit puede admitir (y admite) los dos puntos señalados por Kagan, sin que esto suponga reconocer que su primer argumento fracasa. Kagan equivoca el sentido de este argumento,

"Kagan escribe que, si pasamos de la función maestra a los objetivos ordinarios, S es de nuevo tan tolerante como P, puesto que S no necesita afirmar que alguno de estos objetivos es "racionalmente inaceptable o inferior" (Kagan, p.751). Pero, si estos otros objetivos no son inferiores, ¿por qué el interés propio ha de ser la función maestra? Si esos otros objetivos no son menos racionales, ¿por qué hemos de creer que, cuando entran en conflicto con nuestro propio interés, es irracional actuar según ellos?"²¹.

La confusión de Kagan probablemente obedece al hecho de que no reconoce claramente la distinción entre preferencias morales y preferencias personales, sino que mas bien la confunde con la que existe entre deseos egoísta y no-egoístas. Como muy bien reconoce Kagan, lo que caracteriza a S es que sólo mis preferencias, sean egoístas o no, me dan una razón para actuar. Pero lo que hace a S restrictiva es que sólo admite mis preferencias personales como razones y no mis preferencias morales. Y lo que hace arbitraria esta restricción es el hecho, admitido por T, de que mis preferencias morales pueden ser racionales. Por tanto, como ya dijimos, la cuestión es que, si mis preferencias morales son tan racionales como mis preferencias personales, ¿por qué ha de ser irracional actuar según estas preferencias?

En su crítica, Kagan señala también que el primer argumento de Parfit si bien es una objeción directa (aunque no admisible) para S, sólo proporcionaría (en caso de ser satisfactoria) un apoyo indirecto para P. Esto sucede porque admitir la racionalidad de distintos patrones de pre-

²¹ Parfit (1985), p.845.

ferencias no implica inmediatamente la admisión de P. De hecho, T admite la racionalidad de las preferencias morales. Es decir, T puede admitir (CP2). Esta admisión no es suficiente para que la teoría no sea restrictiva. Parfit reconoce esto, y señala que se necesita un paso más para que el argumento contra T se convierta en un argumento a favor de CP. Este paso es el necesario para hacer que CP sea no-restrictiva. Es decir, CP tendría que afirmar que

(CP3) "si hay varios objetivos que son igualmente racionales, lo que es racional hacer para *mi* debe depender de cuáles, entre estos objetivos, sean *míos*. Y lo que es racional hacer para *mi* ahora debe depender de cuáles, entre estos objetivos, sean *míos* ahora"²².

(CP3) no es más que la explicitación del objetivo de CP, que, como ya vimos, afirma que lo que da a un agente una razón para actuar es aquello que mejor satisface sus deseos racionales presentes. De modo que si suponemos, como en nuestro ejemplo, que tanto la preferencia personal por no confesar como la preferencia moral por confesar son racionales, lo que yo debo hacer racionalmente depende de cuál de estas preferencias sea *mi*a. Si lo que yo deseo es únicamente satisfacer mis preferencias personales, entonces lo racional es satisfacer esta preferencia. Pero si mi mayor deseo es actuar moralmente, entonces lo que es racional es satisfacer mis preferencias morales.

Una vez que aceptamos la racionalidad de las preferencias morales, parece que no aceptar (CP3) sería arbitrario. Cuando el interés propio entra en conflicto con la moralidad, si no podemos señalar ninguna de las dos pautas de conducta como racionalmente suprema, debemos admitir que sería racional seguir cualquiera de ellas. La elección sólo puede depender de *mi*. Lo que sea más racional para *mi*, puesto que no depende del carácter en sí de los distintos patrones de conducta, ha de depender de cuál de ellos sea *mi* objetivo. Es decir, con palabras de Parfit, "¿sería racional para un criminal rendirse a la policía incluso si su único objetivo es escapar a Brasil? y ¿sería racional para él escapar, incluso si su único objetivo fuera rendirse?"²³

Podemos concluir entonces que el primer argumento de Parfit es un

²² Parfit (1986), p. 944.

²³ Parfit (1986), p. 944.

buen argumento. Es decir, debemos rechazar CPS y admitir (CP2) y (CP3). Parfit asume que todas las versiones de S son equivalentes a CPS y deben por tanto ser rechazadas. Sin embargo, nuestra versión crítica T admite (CP2). Pero esto no es suficiente. Sólo si T acepta también (CP3) sería admisible o, al menos, se vería libre del cargo de ser arbitrariamente restrictiva. Volvamos a nuestro ejemplo y supongamos que mi mayor deseo es actuar moralmente. Yo sé que para actuar moralmente debo confesar. También sé que esto va en contra de la satisfacción de mis preferencias personales. Pero no me importa. Aun así, lo que yo más deseo es confesar. ¿Me diría T que no debo hacerlo, que confesar sería irracional? Es decir, ¿afirma T que lo racional es actuar según nuestras preferencias personales racionales o según nuestras preferencias racionales, sean estas personales o morales?

T supone que todos tenemos preferencias personales, y que habitualmente deseamos su satisfacción. Por otro lado, T admite que todos tenemos preferencias morales, en tanto que estas se definen como las preferencias personales hipotéticas que tendríamos si tuviéramos que elegir en una determinada situación de incertidumbre. Pero T afirma que, puesto que realmente no nos encontramos en una situación de ese tipo, no tenemos por qué estar dispuestos a juzgar el mundo desde un punto de vista moral, e.d., que no tenemos por qué desear actuar desde un punto de vista moral. Pero T no afirma que es irracional hacerlo.

Es decir, T no sólo afirma que hay un modo en el que las preferencias morales pueden ser racionales (CP2), sino también que puede ser racional, o al menos que no es irracional, desear actuar desde un punto de vista moral (CP3). T admite la racionalidad de la conducta moral cuando deseamos ser morales. Sin embargo, es posible que T sea en este punto una teoría confusa. T afirma que actuar moralmente es irracional desde el punto de vista del interés personal. Y afirma que, puesto que no estamos en la situación de elección que hemos caracterizado como punto de vista moral, en tanto que agentes racionales, no puede exigiérsenos actuar moralmente. Si yo desear confesar, T me dirá que esto va en contra de mis intereses personales. Pero si yo no desear maximizar mis intereses personales sino satisfacer un determinado criterio de conducta moral, T no puede decirme que soy irracional al confesar, puesto que confesar es la conducta que mejor satisface ese criterio. T no puede criticar mi con-

ducta dados mis objetivos. Lo único que podría hacer es criticar estos objetivos. Pero si T admite que estos son racionales, no puede tampoco criticarlos. Es decir, una vez que T admite (CP2) tiene forzosamente que admitir (CP3). Lo que yo afirmo es que T admite (CP3).

Ha de tenerse en cuenta que T no es una teoría formulada de nuevas sobre el vacío. Es una teoría real que muchos defienden. Y, entre los que defienden T, algunos admiten explícitamente (CP2) y (CP3). Por ejemplo, Harsanyi admite que cuando un agente desea satisfacer un determinado criterio de conducta moral, entonces lo racional para él es actuar moralmente. El problema es que en ocasiones T confunde a) lo que deseamos, e.d., los deseos que, si son racionales, nos dan una razón para actuar, y b) lo que queremos en el sentido de nuestras preferencias personales. Al definir la utilidad como una medida de nuestras preferencias personales racionales, y al afirmar que la conducta racional es la que maximiza esta utilidad, T parece implicar que es irracional no actuar según estas preferencias. Pero afirmaciones como la de Harsanyi nos dan razones para suponer que, si yo no quiero en un momento determinado maximizar mi utilidad, sino satisfacer un criterio de conducta moral, entonces lo racional para mí es hacer esto último. En cualquier caso, si T admite no sólo (CP2) sino también (CP3) entonces escapa a la crítica del primer argumento de Parfit y, en caso contrario, CP es preferible a T al menos en cuanto que T es arbitrariamente restrictiva.

El segundo argumento de Parfit²⁴

El primer argumento de Parfit a muestra que el sesgo a favor de uno mismo no es supremamente racional. Sin embargo, T puede ser aún defendida del siguiente modo. T puede aceptar que el interés propio no es supremamente racional y, aun así, puede seguir siendo una teoría distinta de P. T puede argumentar que la diferencia fundamental que habla en su favor es que la fuerza de la razón se extiende a través del tiempo, e.d., que la razón no es relativa al tiempo. Parfit reconoce esta diferencia, y su segundo argumento está destinado a mostrar que, al igual que ocurría en el caso anterior, S es criticable en este sentido.

La diferencia entre S y P en este aspecto es que mientras que S es parcialmente relativa o híbrida, P es una teoría pura. Diremos que una teoría es pura si, o bien admite tanto la relatividad temporal como la relatividad con respecto al agente, o bien si rechaza ambos tipos de relatividad y es neutral con respecto al tiempo y con respecto al agente. En el primer caso se encuentra P, que afirma que las razones para actuar son relativas al agente en el momento de actuar. Llamaremos a las teorías de este tipo teorías puras relativas. En el otro caso se encuentra, por ejemplo, la teoría de la benevolencia racional de Sidgwick, que mantiene una neutralidad en ambos aspectos. A estas teorías nos referiremos con el nombre de teorías puras neutrales. Por otra parte, diremos que una teoría es híbrida si admite la relatividad en un aspecto pero es neutral en otro. Este sería el caso de S, que mantiene que las razones para actuar son relativas al agente pero temporalmente neutrales. S es por tanto incompletamente relativa, y lo que Parfit intenta mostrar es que una teoría de este tipo puede ser acusada de "un cierto tipo de inconsistencia".

El primer paso dado por Parfit es argumentar que, si las razones pueden ser relativas, entonces pueden ser enteramente relativas. Llamemos a esta afirmación (P1). (P1) apela a la analogía formal existente entre uno mismo y el presente: de la misma manera en que la palabra "yo" se refiere a una persona particular concreta, la palabra "ahora" se re-

²⁴ Parfit no presenta este argumento como independiente del anterior, sino como un argumento a una de las respuestas que los defensores de S pueden dar al primer argumento (ver Parfit, 86, pp.137-148). No obstante, creo que puede ser considerado, tal y como hacemos aquí, como un argumento independiente.

fiere a un momento concreto, el presente. Por ello, si una razón puede ser relativa a mí, puede también ser relativa a mí ahora. Esta analogía formal, según Parfit, es suficiente para exigir que una teoría les de el mismo tratamiento, e.d., para exigir que una teoría sea pura. El segundo paso consiste en afirmar que las razones para actuar deben ser enteramente relativas. Llamemos a esta afirmación (P2). (P2) es la afirmación distintiva de P. A partir de estas afirmaciones, P puede afirmar

(P3) "Consideremos cualquier par de afirmaciones relacionadas del siguiente modo. La primera contiene la palabra "yo" pero no contiene la palabra "ahora". La segunda es exactamente igual que la primera salvo porque también contiene la palabra "ahora". Llamemos *análogas* a estas parejas de afirmaciones. Si la primera de ellas entra en conflicto con la segunda, entonces es incompatible con la relatividad total, y debe por tanto ser rechazada. Si la primera no entra en conflicto con la segunda, entonces es una cuestión abierta si debemos o no aceptarla. Pero si la aceptamos, debemos también aceptar la segunda. Esto es necesario porque, en caso contrario, nuestra teoría sería incompatible con la relatividad total y no estaríamos dando a las expresiones "yo" y "ahora" el mismo tratamiento, cosa que hemos visto que debemos hacer."²²

Las afirmaciones (P1) a (P3) expresan lo que Parfit llama "Apelación a la relatividad plena". Y, obviamente, tal apelación habla a favor de P y en contra de S. Sin embargo, no está en absoluto claro por qué hemos de hacer tal apelación. En efecto, T, al igual que cualquier otra versión de S, es parcialmente relativa. Pero no se ve muy claro por qué esto ha de preocuparnos en absoluto. Realmente ¿cuál es la objeción que puede hacerse a una teoría parcialmente relativa?. Desde luego, una teoría de este tipo es incompatible con la relatividad total. Pero decir esto no representa una crítica a menos que sepamos por qué hemos de aceptar que es deseable que una teoría sea pura.

Parfit afirma que una teoría híbrida presenta "un cierto tipo de inconsistencia". No es, desde luego, inconsistencia lógica. Todo lo que sucede es que T no es constante es su actitud acerca de la relatividad de las razones. Pero ¿por qué habría de serlo? El motivo que parece tener Parfit para considerar exigible una actitud constante hacia la relatividad es la analogía formal existente entre "yo" y "ahora". Podemos aceptar

²² Parfit (198), p.141.

que se da tal analogía formal, pero, tal y como argumenta Kagan²⁶, tal analogía es compatible con la existencia de fundamentos objetivos para tratar estas dos dimensiones del agente y el tiempo de modo distinto.

Sin embargo, Kagan pasa por alto la parte fundamental de la argumentación de Parfit. Lo que hace deseable que una teoría sea pura es el hecho de que una teoría híbrida puede ser atacada por dos direcciones distintas. Por un lado, debe enfrentarse con las teorías puras relativas y, por otro con las teorías puras neutrales. Y todo movimiento de defensa por parte de T ante el ataque de una de ellas puede volverse contra ella frente al ataque de la otra.

Supongamos por ejemplo una teoría moral consecuencialista que afirme que debe darse el mismo peso a los interés de cualquiera en cualquier momento. Por ejemplo, Parfit cita la pregunta de Rashdall "¿por qué una Razón imparcial e impersonal ... ha de dar más importancia al placer de un individuo que al de otro?"²⁷ Un defensor de T podría contestar diciendo que la pregunta no es qué es lo que es racional hacer para "la Razón", sino que es racional hacer para *mi*. Desde, luego, si adoptamos un punto de vista de la totalidad, este punto de vista a de ser imparcial. Pero ¿por qué habría de adoptarlo? Después de todo, yo no soy la totalidad. Yo soy yo. Y por ello, las razones que han de ofrecerse han de estar referidas a mí. Esta defensa de T puede ser válida contra el ataque del consecuencialismo. Pero, de modo similar, P puede atacar a S diciendo que la pregunta no es qué es lo racional para mí, sino qué es lo racional para *mi ahora*.

Es esta vulnerabilidad la que hace deseable una teoría pura. Y esta vulnerabilidad surge de la analogía entre "yo" y "ahora". Es cierto que la decisión he de tomarla yo. Pero es igualmente cierto que he de tomarla *ahora*. En contra de la relatividad temporal, T puede argumentar que, después de todo, las decisiones que yo tomo ahora me afectarán en el futuro. Pero el consecuencialismo puede contestar que no es menos cierto que las decisiones que yo tomo afectan a otras personas.

²⁶ Kagan (86), p.753.

²⁷ Parfit (86), p.142.

Sin embargo, la vulnerabilidad de T en este sentido no tiene por qué resultar un problema grave. Puede que T sea capaz de responder a las acusaciones de ambos frentes. Su defensa, desde luego, ha de consistir en argumentar que la pregunta pertinente es realmente la pregunta acerca de lo que es racional para mí, y no la de qué es lo racional para la Razón imparcial ni tampoco la de qué es racional para mí ahora. Y la crítica a T sólo puede ser decisiva si disponemos de una razón para mantener que esa no es la pregunta fundamental. Parfit no ofrece ninguna razón, al menos en este momento. Se limita a constatar la vulnerabilidad de T. Pero esto no es un problema, en tanto que T sea capaz de contestar a) por qué yo no tengo motivos para preocuparme de los intereses de los demás y b) por qué sí tengo motivos para preocuparme de mis intereses futuros.

En lo que sí tiene éxito el segundo argumento de Parfit es en dejar a T la carga de la prueba. Es un hecho que yo tengo que decidir ahora. Parece por tanto que debo tener en cuenta mis interés presentes. Es decir, lo que sí es cierto es que, una vez que admitimos que las razones pueden ser relativas, entonces pueden ser totalmente relativas. Debemos por tanto aceptar (P1). Y también es cierto que si admitimos (P1), y sin embargo queremos mantener que las razones deben ser temporalmente neutrales, debemos explicar qué es lo que, mas allá de la analogía formal entre "yo" y "ahora", diferencia ambas dimensiones.

Podemos concluir el análisis del segundo argumento de Parfit con las siguientes conclusiones. T es una teoría híbrida, mientras que P es una teoría pura. Por ser híbrida, T no es inconsistente ni incoherente. Sin embargo, su carácter híbrido hace de T una teoría extremadamente vulnerable, pues puede ser atacada por dos frentes. Sin embargo, esto no hace que T haya de ser rechazada, al menos si puede salir airosa de estos ataques. Podemos incluso admitir que es deseable que una teoría sea pura. Pero que sea deseable no quiere decir que sea posible. (P1) afirma que es posible una teoría pura relativa. Esto es cierto en un sentido, en el cual también es posible una teoría pura neutral, a saber, en el sentido de que una teoría pura puede ser descrita coherentemente. Pero (P1) no muestra que tal teoría pueda ser satisfactoria. T puede asumir (P1) y afirmar que una teoría pura relativa, aunque coherente, sería falsa y objetable. Puesto que el argumento de Parfit se dirige únicamente al carácter vulnerable de T y no a si T puede o no puede contestar de forma

satisfactoria a los ataques, dejaremos este último asunto para mas adelante, y pasaremos ahora a analizar el último argumento de Parfit..

El tercer argumento de Parfit

Este tercer argumento intenta contestar, al menos en parte, la cuestión substantiva de si S puede justificar de algún modo su carácter parcialmente relativo frente a los ataques de P. Es decir, se trata ahora de preguntarse si es justificable la neutralidad temporal. Parfit analiza esta cuestión partiendo de una versión concreta de S, a saber, de la versión que conocemos como (DF), según la cual lo que una persona tiene mas razón para hacer es aquello que se espera que satisfaga mejor el conjunto de sus deseos. Es importante recordar que es esta versión de S la que critica Parfit pues, al contrario de lo que sucedía con el segundo argumento, esta crítica no es extensible a otras versiones de S.

El procedimiento que sigue Parfit consiste en mostrar las conclusiones absurdas que se siguen de una teoría que, como (DF), pretenda dar el mismo peso a todos los deseos del agente, sean estos pasados, presentes y futuros. La consecuencia indeseable sería que, según (DF), yo debo darle el mismo peso a un deseo que tuve en el pasado, pero que ya no tengo, que a uno que tengo actualmente.

Esto es cierto al menos de los deseos no-condicionales. Cuando un deseo es condicional en su propia persistencia, puedo desecharlo una vez que está en el pasado. Si por ejemplo, yo quería disfrazarme durante el carnaval, mi deseo estaba condicionado en su persistencia a la duración del carnaval. Por ello, ahora que los carnavales han pasado, yo puedo desechar ese deseo pasado. Hay un tipo de deseos que siempre son implícitamente condicionales en este sentido, a saber, aquellos deseos cuya satisfacción suponemos que nos causará placer o cuya no satisfacción suponemos que nos causará infelicidad. Estos deseos se suponen condicionados al momento en que su satisfacción o insatisfacción tendrá estos efectos. Y supuesto que este tipo de deseos, una vez que ya no los tenemos, no pueden afectarnos de ese modo, no tenemos por que darles ningún peso. Este es el motivo por el que la versión hedonista de S (H) no implica que debamos darle peso a los deseos pasados, pues por definición, todos los deseos relevantes desde el punto de vista de (H) son implícitamente condicionales en este sentido. Sin embargo, según (DF) los deseos nos dan una razón para actuar no en tanto que su satisfacción nos produce placer y su insatisfacción infelicidad, sino simplemente por el hecho de

ser deseos. Y puesto que hay deseos incondicionales, cuando los deseos pasados son de este tipo debemos darles el mismo peso que damos a los deseos presentes y futuros. Un ejemplo de deseo incondicional sería el siguiente

"Supongamos que, durante quince años, yo no sólo he trabajado intentando salvar a Venecia, sino que incluso he cotizado regularmente a la Fundación para la preservación de Venecia. A lo largo de esos quince años mi más fuerte deseo ha sido que Venecia fuera salvada y que yo fuera uno de sus salvadores. Esos deseos no eran condicionales en su propia persistencia. Yo podría preguntar ¿quiero que Venecia sea salvada, y que yo sea uno de sus salvadores, incluso si más adelante dejo de tener tales deseos? Mi respuesta sería SI."²⁸

Dejando aparte los deseos condicionales, (DF) afirma que debemos dar el mismo peso a todos nuestros deseos, ocurran estos en el momento que ocurran. Pero entonces, esto significa que, aun en el caso de que deje de desear que Venecia sea salvada, incluso si empiezo a pensar que no merece la pena hacer ningún esfuerzo para salvar algo tan insalubre, debo, según (DF) seguir cotizando a la Fundación. Es mas, según (DF) la mejor para mí es que Venecia sea salvada, a pesar de que yo ya no lo deseo en absoluto, a menos que durante más de quince años desee que no se salve con la misma fuerza con la que antes deseaba su salvación, o durante menos tiempo si mi deseo de que no se salve es suficientemente más fuerte que mi anterior deseo de salvarla. (DF) afirmará que sería irracional por mi parte no contribuir a la salvación de Venecia incluso cuando ya no tengo el mas mínimo deseo de salvarla.

Esta conclusión sin duda es absurda y resulta extremadamente embarazosa para (DF). Supongamos que, para evadir esta conclusión, (DF) afirma que no debemos dar igual peso a todos nuestros deseos, sino sólo a los deseos presentes y futuros. Esta pretensión, aunque defendible, ya no puede ser sostenida apelando al requisito de neutralidad temporal (NT en adelante). Si no debemos dar peso, o al menos no el mismo peso, a los deseos pasados *porque* son pasados, entonces es que (NT) no es el motivo para dar el mismo peso a los deseos presentes y futuros, sencillamente *porque* no podemos defender (NT), ya que su defensa implicaría que también debe darse el mismo peso a los deseos pasados. Pero entonces, ¿cómo

²⁸ Parfit, (198), p.152.

puede defender (DF) su afirmación de que debe darse el mismo peso a los deseos presentes y a los futuros?

Esta afirmación se vuelve ahora extremadamente difícil de defender. Si los deseos futuros tampoco son mis deseos ahora, ¿por qué he de darles algún peso? No puede contestarse que "serán *míos* alguna vez", puesto que, incluso si sabemos con certeza que tendré esos deseos, también sé con certeza que los deseos pasados fueron *míos*. Por tanto, es posible que (DF) prefiera, a pesar de todo, sostener (NT). Pero entonces (DF) debe dar otro motivo para desechar los deseos pasados que no sea al mismo tiempo un motivo para desechar los futuros. Por ejemplo, (DF) puede apelar a los motivos por los que cambian los deseos y afirmar que podemos desechar los deseos que se han perdido a causa de que el agente ha cambiado de opinión.

Esta respuesta no es muy satisfactoria. En primer lugar, no cubriría todos los deseos, sino sólo aquellos que se basan en juicios de valor. Muchos de mis deseos no dependen de juicios de valor en este sentido. Es cierto que todos los deseos, descontando los implícitamente condicionales (que abarcarían lo que nosotros hemos llamado "preferencias primarias") implican algún tipo de valoración. Pero, en muchos casos, el cambio de deseos no implica un cambio de valoración. Hay muchas cosas que yo considero valiosas, entre otras ser un gran atleta y batir un récord mundial, ser una gran bailarina, hacer un gran descubrimiento científico, ayudar a aliviar la miseria del tercer mundo, ver mundo y dedicar la vida a la contemplación. Pero dadas nuestras limitaciones, yo no siento el mismo deseo por todas ellas, a pesar de que puedan parecerme igualmente valiosas. Supongamos que hace diez años mi mayor deseo era ser una gran bailarina. Mi deseo ha cambiado y ahora lo que más deseo es dedicarme a la filosofía. Este cambio no implica un cambio de opinión. Yo no valoro ahora menos el baile ni valoraba antes menos la filosofía. Simplemente, antes deseaba una cosa y ahora otra²⁹. En estos casos, la pregunta sigue sin contestación.

Para estos deseos, (DF) puede afirmar, o bien que debo dar el mis-

²⁹ Puede pensarse que el anterior deseo de ser bailarina era un deseo condicional, cuya auténtica expresión era "deseo ser bailarina mientras aun tengo edad para ser una primera figura". Esto no es cierto. Mi deseo no era condicional. De cualquier modo, puede sustituirse este deseo por el de ser un gran poeta. Nunca es tarde para la poesía.

mo peso a los deseos futuros por el requisito de neutralidad temporal, con lo cual deberá también dar peso a los pasados, lo cual resulta muy insatisfactorio, o rechazar (NT) y tratar de buscar otro motivo que discrimine entre deseos pasados y futuros, cosa que, como ya dijimos, resulta sumamente difícil. Pero incluso para los deseos que si dependen de juicios de valor, esta respuesta, aunque plausible, no serviría para apoyar (DF) sino más bien todo lo contrario, pues lo mismo que puede argumentarse con respecto a los deseos pasados puede decirse respecto a los deseos futuros. Mis deseos futuros, si no son los mismos de ahora, también se fundamentarán en un cambio de opinión. Si mis deseos futuros se basarán en unos juicios de valor que yo ahora rechazo, al igual que los pasados se basaban en otros que también son rechazados en el presente, entonces ¿qué motivo puedo tener para no dar el mismo peso a los pasados y dárselo a los futuros?

No parece que haya una salida buena para (DF). Sin embargo, como hemos dicho, la apelación a los motivos por los que cambia un deseo es plausible. Lo que sucede es que (DF), por su propio carácter, no puede utilizar esta apelación. (DF) trata los deseos como "meras preferencias", es decir, como algo que un individuo tiene de hecho o no tiene. Incluso cuando los deseos y los cambios de deseos se basan en algún tipo de juicio de valor, (DF) no dispone de ningún mecanismo para discriminar entre ellos ni puede afirmar que la satisfacción de algunos es racional mientras que la de otros no lo es.

Esto sucede porque (DF) no es una teoría crítica. Y por eso tiene que dar el mismo peso a mis deseos pasados presentes o futuros, pues todos ellos son igualmente mis deseos. O, si quiere desechar los pasados porque son pasados, debe dar un motivo, nada fácil de encontrar una vez que se rechaza (NT), para no desechar los futuros porque son futuros. Y, sin rechazar (NT), (DF) no puede acudir a los cambios de juicios de valor, ya que, supuesto que todos ellos son igualmente válidos, sólo pueden distinguirse entre sí por el momento de su ocurrencia temporal. Y, de nuevo, si rechazamos las opiniones pasadas porque son pasados, debemos igualmente desechar las futuras porque son futuras.

Una teoría que no entienda los distintos deseos como meras preferencias puede evitar este problema. Si no lo son, entonces algunos de mis deseos son racionales y otros no, y esto depende básicamente de que estén

basados o no en juicios de valor justificables. En ese caso tendremos un motivo para tratar de modo distinto a deseos distintos y dar a unos más peso que a otros. Tomemos un ejemplo de Nagel citado por Parfit³⁰, especialmente bueno porque refleja algo que le pasa a casi todo el mundo. Supongamos que yo ahora valoro ante todo el sexo, la espontaneidad, el riesgo y las emociones fuertes. Y que sé que dentro de veinte años lo que valoraré por encima de todo será la seguridad, el status, el dinero y la tranquilidad. La cuestión es qué debo hacer ahora, a que deseos debo ahora darles peso.

Habitualmente, este ejemplo no representa un problema, ya que por lo general tales deseos son implícitamente condicionales. Yo deseo todo eso ahora que tengo veintitantos años, pero sé que todo esto está ligado a mi edad y que lo quiero ahora. Después querré otra cosa y lo anterior me parecerá tan poco valioso como me lo parece ahora la preocupación mezquina por el dinero y el afán pequeño-burgués por la seguridad. Pero supongamos que estos deseos son incondicionales, tal y como algunas veces, en especial en la adolescencia, nos parecen, y que podemos decir que deseamos vivir de modo bohemio incluso cuando se nos haya curado el deseo. Si todo son preferencias, entonces parece que inevitablemente debo ponerle una vela a Dios y otra al diablo y dar el mismo peso a todos mis deseos.

Pero supongamos que estoy convencida de que lo que me sucederá con el paso de los años puede ser considerado como una degeneración del carácter que acompaña inevitablemente a la decadencia física. O bien que creo firmemente que mi sangre ahora es más fuerte que mi cerebro y que sólo con los años adquiriré la carga de sabiduría necesaria para juzgar el mundo. Y supongamos que, sea cual sea el caso, tengo buenos argumentos en los que apoyarme. Entonces, si mi opinión es la primera, tengo buenos motivos para no dar el mismo peso a los deseos futuros. Es más, no querré hacer nada que pueda favorecer al cerdo repugnante en el que voy a convertirme sin remedio. O, si es la segunda, tengo buenos motivos para no dar peso, o al menos demasiado peso, a mis deseos presentes, pues intentaré sujetar lo mejor posible lo que yo entiendo que es producto de la enfermedad de la juventud, esperando con paciencia el remedio del tiempo.

³⁰ Parfit, (1961), p.154.

En cualquier caso, lo mejor que puedo hacer es actuar según las preferencias que me parecen válidas, o al menos de las que ahora me lo parecen. Tengo un motivo para discriminar entre los distintos deseos. Y este motivo está ligado al tiempo. Me hace desechar los deseos futuros o los presentes. Pero no porque sean futuros, sino porque en el futuro serán peores.

Podemos de este modo admitir (NT) y aun así discriminar entre deseos pasados, presentes y futuros. Podemos mantener que es irracional dar menos peso al futuro por ser futuro, y al mismo tiempo mantener que debo darle menor peso a mis deseos futuros porque el tiempo es corrupción. A partir de esto, Parfit concluye

"Según S, debo dar el mismo peso a todos mis deseos presentes y futuros. Esta afirmación se aplica incluso a aquellos deseos futuros que dependerán de un cambio en mis juicios de valor o mis ideales. Cuando se aplica a estos deseos, la afirmación es indefendible. En el caso de las razones para actuar que se basan en juicios de valor o ideales, un agente racional debe dar prioridad a los valores o ideales que él acepta ahora. En estos casos, la teoría correcta no es *S* sino *P*."⁵¹

Sin embargo, esta conclusión es cuestionable, pues es discutible que ninguna versión de S pueda dar cuenta de este hecho. En concreto, es discutible que T no pueda hacerlo. T es una teoría crítica y, por tanto, no trata los juicios de valor como meras preferencias o, dicho más claramente, no trata todas las preferencias de igual modo, sino que, al contrario, afirma que hay preferencias y conjuntos de preferencias que cumplen determinadas exigencias de racionalidad y otros que no las cumplen. Supongamos que mis preferencias presentes no son racionales. Entonces T nos dice que no hemos de darles peso en nuestras consideraciones. O supongamos que podemos predecir que nuestras preferencias futuras serán irracionales. Entonces T nos dirá que no debemos darles peso y que debemos, por el contrario, tomar medidas para prevenir nuestra actuación según estas preferencias irracionales cuando estas se hagan presentes. Este tipo de casos son precisamente los que analizábamos en el capítulo anterior al hablar de la racionalidad indirecta.

Esto no significa que T y P sean equivalentes. P, o al menos CP, afirma que, puesto que los juicios de valor en los que se apoyan deter-

⁵¹ Parfit (1984), p.155.

minadas preferencias, pueden ser más o menos justificables, no debemos darles a todos el mismo peso, sino que debemos dar más peso a aquellas preferencias basadas en juicios de valor que nos parecen correctos ahora. T afirma lo mismo. Pero la cuestión de fondo es por qué se afirma esto. Hay dos alternativas. O bien debemos darle peso a las preferencias apoyadas en juicios de valor que nos parecen correctos ahora, por el hecho de ser ahora cuando nos lo parecen, o bien debemos darle peso a las preferencias basadas en juicios de valor que nos parecen correctos. Naturalmente, tenemos ahora los juicios de valor que nos parecen correctos, pero esto sólo trivialmente es una aceptación de una teoría no neutral temporalmente. Lo que hace válidos a ciertos juicios de valor no es que los mantengamos ahora, sino una serie de motivos que, naturalmente, si los tenemos, los tenemos ahora. No se puede tener un motivo, ni conocer un argumento, y no tenerlo ni conocerlo ahora. Esto es una verdad de Perogrullo. Tal y como se deduce de los textos de Parfit, el argumento válido es este último. Pero este argumento no es un argumento a favor de P, sino a favor de T. P sería defendible si el argumento fuera el primero, y los juicios de valor que deben prevalecer son los que nos parecen correctos ahora debido a su ocurrencia temporal en el presente. Si este argumento es indefendible, entonces no hay aquí nada que hable en favor de P.

Tal y como ya ha ocurrido otras veces, Parfit parece confundido porque no admite ninguna versión crítica de S. Como el argumento relativo a los juicios de valor habla en favor de una teoría crítica, y Parfit sólo tiene en cuenta la versión crítica de P, piensa que este argumento favorece a P. Pero el argumento a favor de CP, la versión crítica de P, no apoya esta versión por ser una versión de P, sino por ser crítica. Por ello mismo, este argumento también habla a favor de T, la versión crítica de S. El tercer argumento, por tanto, no apoya P en contra de S, sino las versiones críticas contra las versiones no críticas.

Es más, del propio argumento de Parfit se siguen razones a favor de T y en contra de CP, salvo que CP se haga equivalente a T. Pensemos en las preferencias cuyo cambio no significa un cambio de juicio de valor. En el ejemplo de Nagel, supongamos que tanto la vida tranquila y segura como la bohemia me parecen igualmente valiosas. Yo ahora prefiero la vida bohemia y sé que en un futuro preferiré la seguridad. Y ninguna de estas preferencias es condicional. Entonces, lo que T afirma es que debo dar

igual peso a mis preferencias presentes y a las futuras. Sé que un exceso de bohemia en el presente impedirá que mi preferencia futura por la vida tranquila se satisfaga. Si esto sucede, el balance total de utilidad a lo largo de mi vida será menor. Y sé que si ahora me modero un poco, cuando cambien mis preferencias aun podré satisfacerlas suficientemente. Cuánto he de moderarme depende de la fuerza relativa de estas preferencias. Incluso ahora, puede argumentarse que hay razones para preferir las preferencias actuales. Pero el motivo no es que sean actuales, sino que ahora sé con qué fuerza deseo vivir de un modo bohemio mientras que no puedo calcular con exactitud cuánto preferiré en el futuro la vida tranquila. Pero si lo supiera, entonces el cálculo dependería solamente de esta dato.

Si P es una teoría distinta de T, tiene^{que} afirmar que en estos casos debemos darle más peso a las preferencias presentes porque son presentes. Y no parece haber ningún motivo que justifique esta afirmación. Y si se admite la afirmación de T, entonces ambas teorías serían iguales en este caso, pero no porque T acoja las afirmaciones de P, sino porque P hace suyas las de T. Es decir, llamemos como llamemos a la teoría que realice estas afirmaciones, esta teoría será T.

El motivo para dar más peso a las preferencias presentes nunca puede ser el hecho de que sean presentes. Cuando esta es la única diferencia entre las preferencias, T exige que se les dé el mismo peso. Esto significa que T es temporalmente neutral. No obstante, parece que aún hay algo a favor de CP o, al menos en contra de T, a saber, lo concerniente a las preferencias pasadas. Pues, precisamente por ser temporalmente neutral, parece que T ha de dar también peso a las preferencias pasadas no condicionales. Esto no sucedería en la versión hedonista de S³². En esta versión, lo que da a un agente una razón para actuar es la satisfacción de sus deseos personales en tanto que esta satisfacción le produce placer. Pero puesto que la satisfacción de los deseos pasados no produce placer, no tenemos por qué darles ningún peso.

Esta versión de S es, tal y como argumenta Parfit³³, inaceptable.

³² Parfit, (1961), p.149-150.

³³ Parfit (1961), p.493.

Sin embargo, hay una variante de esta versión de S, a la que Parfit llama *Preference-Hedonism*, que no está sujeta a esas críticas y de la cual, como ya dijimos en su momento³⁴, T sólo se distingue en su carácter crítico. Y en esta variante más aceptable podemos también afirmar que no debe darse peso a las preferencias pasadas en tanto que su satisfacción no significa un incremento de la utilidad. Es decir, esta variante, y T en tanto que es su versión crítica, escapa a la consecuencia absurda de tener que afirmar que debe dársele peso a unas preferencias en cuya satisfacción ya no estamos interesados, sin por ello renunciar a la defensa de la neutralidad temporal característica de todas las versiones de S³⁵.

Hemos revisado los tres argumentos que, analizando las diferencias entre S y P, hablan según Parfit en contra de S y a favor de P. El resultado de esta revisión puede resumirse del siguiente modo. El primer argumento ofrece la crítica más fuerte. En ella, Parfit muestra que el sesgo a favor de uno mismo no es supremamente racional. Sin embargo, nosotros hemos defendido que T no muestra tal sesgo. En cualquier caso, si lo hace, T es rechazable bajo el cargo de ser arbitrariamente restrictiva. El segundo argumento, sin embargo, no constituye una crítica en absoluto. Todo lo que se muestra es que T, por ser una teoría híbrida, que exhibe una relatividad parcial, es atacable por dos frentes, lo que la hace más vulnerable que una teoría pura. Pero esto no habla en contra de T a menos que pueda mostrarse que T no puede salir airosa de estos ataques. Por último, el tercer argumento tampoco constituye una crítica a T, sino, en primer lugar, a las versiones no críticas tanto de S como de P

³⁴ Ver p.218, nota 5.

³⁵ En la sección 64 de *Reasons and Persons*, Parfit utiliza algunos ejemplos en los que, al hablar en términos de 'dolor' parece involucrar a las versiones hedonistas de S. Sin embargo, tales ejemplos no suponen una crítica a estas versiones. Si yo no sé si he sufrido una operación muy dolorosa ayer o si voy a sufrir una menos dolorosa ahora, mi preferencia por la primera alternativa no refleja un rechazo del requisito de neutralidad temporal, sino el hecho de que la satisfacción del deseo 'no haber sido operado ayer' no puede alterar en modo alguno el nivel de utilidad, mientras que la satisfacción de la preferencia 'no ser operado ahora' sí puede hacerlo. La pregunta por la utilidad sólo tiene sentido con respecto al presente y al futuro. Una violación del requisito de neutralidad temporal sucedería si, enfrentado con la alternativa de ambas operaciones eligiera la más dolorosa sólo por que es más lejana. En un sentido relevante, el pasado no existe (e.d., no existe directamente, sino por sus secuelas). Parfit puede contestar que el futuro tampoco existe. Esto es cierto, pero también lo es que alguna vez existirá. Prefiero haber tenido una infancia dolorosa y un futuro espléndido a haber sufrido poco antes y que me espere un sufrimiento considerable. Pero esto sucede sólo porque el pasado ya no me afecta ni me afectará nunca más. De hecho, si volviera a nacer, mi preferencia se invertiría.

y, en segundo lugar, a las versiones no hedonistas de S. Es decir, todo lo que puede deducirse de este tercer argumento es que a) las versiones no críticas son inadmisibles y b) que la versión (DF) de S está basada en una mala teoría acerca de lo que es el interés propio.

El tercer argumento tiene una gran importancia indirecta. Lo hemos rechazado afirmando que T es una versión crítica de la variante hedonista de S llamada *Preference-Hedonism*. Pero entonces esto parece mostrar que T debe defender el sesgo a favor de uno mismo, lo que la haría criticable. Es decir, parece que no es posible escapar al mismo tiempo de la crítica del primer argumento y de la del tercero. T escapa a la crítica del primer argumento sólo si admite que no es irracional obrar según las preferencias morales, siempre y cuando esas sean nuestras preferencias más fuertes. Pero "más fuerte" no puede significar "cuya satisfacción rinde mayor utilidad" pues "utilidad" está definida sobre las preferencias personales. Lo único que puede significar es "cuya satisfacción nos parece más importante" o "cuya satisfacción constituye nuestro mayor deseo". Pero entonces ¿puede ser T una versión crítica del *Preference-Hedonism* (PH)?

La respuesta correcta es sí y no. La versión no crítica de (PF) se distingue de la versión hedonista en que (PF) supone que una experiencia es más placentera si es más preferida, e.d., ser más placentero se identifica con tener mayor utilidad. Esto parece servir para las preferencias personales, pero no para las preferencias morales. En este sentido, por muy crítica que sea nuestra teoría, no puede ser una versión de (PF). Pero sí lo es en el sentido relevante para escapar a la crítica del tercer argumento. El objetivo propuesto por T, a saber, la satisfacción de nuestras preferencias racionales, es equivalente, no a la satisfacción de todos nuestros deseos (lo cual la separa de la criticable (DF), sino sólo de aquellos de nuestros deseos cuyo cumplimiento nos interesa (aunque se separe de (PH) en que la expresión "cuya satisfacción nos interesa" no equivale a "cuya satisfacción aumenta nuestra utilidad"). Pero esto es suficiente para justificar que no se dé peso a las preferencias pasadas sin por ello tener que renunciar al requisito de neutralidad temporal, puesto que la satisfacción de nuestras preferencias pasadas, e.d., de aquellas que ya no tenemos, no nos interesa. Es decir, las preferencias relevantes para T son también (y en esto se parece a las dos variantes

hedonistas) implícitamente condicional.

Como hemos dicho, suponemos que T admite que la satisfacción de las preferencias morales no es menos racional que la satisfacción de las preferencias morales. Pero, en cualquier caso, como también hemos dicho, una teoría aceptable acerca de la racionalidad práctica debe admitir esto, lo que supone aceptar la parte más importante de la argumentación de Parfit. ¿Cuál sería entonces la diferencia entre T y CP?. Si no es que T presente el sesgo a favor de uno mismo, y tampoco lo es la neutralidad temporal, entonces, al menos si CP afirma CP1³⁴, parece que es indiferente que adoptemos una u otra teoría.

³⁴ Ver p. 222.

IV.

El hecho de que S sea, en cualquiera de sus versiones, directamente auto-frustrante a nivel colectivo, si bien no hace que S falle en sus propios términos, puesto que se trata de una teoría acerca de la racionalidad individual y no de la racionalidad colectiva, sí hace que S sea en cierto sentido poco satisfactoria. Parfit comparte este punto de vista y, por ello, trata de criticar a S desde una teoría sobre la racionalidad individual alternativa. Sin embargo, sus argumentos, si bien sirven para criticar a algunas de las versiones de S, no resultan válidos frente a nuestra versión crítica T. Al no tomar en consideración las posibles versiones críticas de S, Parfit cree que a) sus argumentos señalan diferencias entre S y P (o al menos entre S y CP) y que b) estas diferencias hablan a favor de CP. Pero, entre nuestra versión crítica de S, T, y la teoría alternativa P, en concreto su versión crítica CP, las diferencias se reducen considerablemente. Esto no es sorprendente desde el momento en que los argumentos de Parfit, en general, mas que señalar diferencias entre S y P, señalan diferencias entre teorías críticas y no críticas. Puesto que tanto T como CP son teorías críticas, su coincidencia es considerable. Y, tal y como hemos visto, cuando ambas teorías se diferencian, la diferencia está a favor de T, y CP sólo puede resultar aceptable si se hace equivalente a T.

Pero aún queda un punto importante. Si la necesidad de encontrar una alternativa a S surge fundamentalmente del hecho de que S sea, como indudablemente lo es, insatisfactoria en cierto sentido, debemos esperar que la alternativa P sea satisfactoria donde S no lo es, es decir, debemos esperar que P, o al menos su versión crítica CP no sea directamente auto-frustrante a nivel colectivo³⁷. T lo es porque, en determinadas situaciones de interacción, da lugar a dilemas. En ellos, todos ganarían más si todos escogieran la estrategia A. Sin embargo, lo mejor para cada uno es escoger la estrategia B, por lo que T afirma que un agente racional debe actuar según esta estrategia. Supongamos una de estas situaciones y preguntémosnos que diría CP. Sea una situación representada por la

³⁷ Es importante notar que este punto no depende de que S y P sean tan distintas como Parfit piensa o tan parecidas como yo he pretendido demostrar. Y no depende de esto porque en nuestra argumentación no hemos disminuido las distancias entre T y CP mediante modificaciones a CP, sino, simplemente, considerando una versión de S a la que Parfit no tiene en cuenta.

siguiente matriz de pagos

	B_1	B_2
A_1	(1,1)	(3,0)
A_2	(0,3)	(2,2)

Ejemplo 12

CP afirma que un agente racional deberá actuar según la estrategia que mejor satisfaga aquellos de sus deseos presentes que no son irracionales. Supongamos que las utilidades representadas son una medida sobre el conjunto de las preferencias presentes no irracionales de los agentes. Por tanto, CP aconsejara a los agentes A y B escoger las estrategias A_1 y B_1 respectivamente. Si esto es lo que sucede, entonces CP también da lugar a dilemas.

Sin embargo, un defensor de CP podría decir: "sí, pero la diferencia real entre S y CP se ve cuando consideramos qué dicen cada una de estas teorías respecto a las estrategias A_2 y B_2 . Para S, escoger estas estrategias es irracional, incluso si el mayor deseo de los agentes es actuar según ellas. Por eso, S nunca puede salir de los dilemas, pues para ello S debería aconsejar una actuación que considera irracional. Sin embargo, para CP no es irracional escoger estas estrategias si ese es el mayor deseo del agente. Por tanto, para CP el dilema tiene una salida". Tal y como hemos dicho, para T tampoco es irracional escoger la estrategia altruista si este es el mayor deseo del agente, pero esto no es lo que interesa ahora.

Sea lo que sea lo que diga T, lo que importa es saber si una teoría que permita como racional elegir estrategias altruistas escapa a los dilemas. Los agentes de nuestro ejemplo pueden estar en una de estas dos situaciones. O bien su mayor deseo es actuar según la estrategia que mejor satisfaga sus preferencias morales, e.d., las estrategias altruistas A_2 y B_2 respectivamente, o bien no siente este deseo o, al menos, este no es su deseo más fuerte. Supongamos que nuestros agentes se encuentran en el primer caso. Si esto es lo que sucede, entonces nuestra situación deja de representar un dilema. Ambos elegirán la estrategia altruista con lo cual a) ambos harán lo racional y b) ambos conseguirán el mejor resultado posible en términos de interés propio.

Pero supongamos que nuestros agentes se encuentran en el segundo caso. Entonces se producirá un dilema pues a) ambos actuarán racionalmente, intentando conseguir la mayor satisfacción de sus preferencias personales y b) sin embargo, y precisamente por actuar de un modo racional, el resultado será subóptimo, e.d., será peor que otro resultado posible en el que los intereses de ambos se satisfacen mejor. Es decir, si esto es lo que sucede, entonces CP será directamente auto-frustrante a nivel colectivo, pues si ambos siguen satisfactoriamente CP, entonces, y a causa de ello, sus objetivos dados por CP se alcanzan peor de lo que se hubieran alcanzado si ninguno de ellos hubiera seguido con éxito CP.

Esto significa que CP no resulta satisfactoria donde T resulta insatisfactoria. Más bien al contrario, ambas teorías conducen a situaciones de dilema y son por ello directamente auto-frustrantes a nivel colectivo. Y esto por la siguiente razón. Cuando los agentes desean actuar desde un punto de vista moral, entonces, como ya sabemos, el dilema desaparece. Pero cuando no lo desean el dilema se presenta.

Para Parfit, la cuestión es si una teoría considera como irracional elegir la estrategia altruista, e.d., actuar desde el punto de vista moral, y por ello cree que una teoría satisfactoria debe considerar la posibilidad de que actuar según criterios morales sea racional. Sin embargo, el problema no es este. Desde luego, una teoría que considerara irracional la elección altruista imposibilitaría la resolución teórica de los dilemas, y sería por ello en algún sentido más insatisfactoria. T, según defendemos aquí, no es una teoría de este tipo. Pero el problema real es si una teoría evita los dilemas, y esto sólo puede hacerlo una teoría que afirme que, al menos en estas situaciones, lo racional es escoger la estrategia altruista, e.d., una teoría que afirme que, en estas ocasiones, la única elección racional es la elección desde el punto de vista moral. Sólo una teoría de este tipo no sería en ninguna ocasión directamente auto-frustrante a nivel colectivo.

Por lo que hasta ahora sabemos, T no es una teoría de este tipo. Ni tampoco lo es CP. Sin embargo, ambas teorías afirman que hay ciertas preferencias, ciertos deseos, que están racionalmente exigidos, en el sentido de que un conjunto de deseos, por el hecho de contener ciertos

elementos, debe también contener algunos otros³⁸. Como ambas teorías reelaboran esta afirmación, podemos analizar la cuestión sólo en una de ellas, por ejemplo en T, y hacer a la otra extensivo el resultado del análisis. Sin embargo, antes de empezar conviene señalar que Parfit deja la cuestión sin resolver, limitándose a afirmar que es posible una versión de CP en la que se afirma

(CP4) "Cada uno de nosotros está racionalmente requerido a preocuparse por la moralidad, a preocuparse por las necesidades de los demás. Puesto que esto es así, tenemos una razón para actuar moralmente, incluso si no tenemos ningún deseo de hacerlo. Que tengamos o no una razón para actuar depende habitualmente de que tengamos o no ciertos deseos. Pero esto no sucede en el caso de los deseos que están racionalmente exigidos"³⁹.

Naturalmente, si CP afirma (CP2) entonces no da lugar a dilemas. Pero la cuestión es si es aceptable o no esta afirmación. Parfit reconoce expresamente que esta es una cuestión polémica y, por tanto, prefiere dejar abierta la cuestión. Sin duda, el motivo fundamental para dejar esta cuestión sin resolver es que, para Parfit, basta con que una teoría no afirme que la conducta moral es irracional. Pero esto, como hemos visto, no es así. Por tanto, tenemos un buen motivo para tratar de resolver esta cuestión. Parfit afirma que para CP hay deseos racionalmente exigidos, afirmación que como ya sabemos también realiza T, pero no da ninguna explicación de cómo es posible que suceda esto, ocupándose tan sólo del caso de los deseos irracionales. Sin embargo, un análisis de los motivos por los que una preferencia puede ser irracional nos da, tal como ya vimos en su momento⁴⁰, una pista acerca de cómo es posible que un deseo esté racionalmente exigido. Veamos esto con un ejemplo de Parfit. Supongamos que un agente muestra un gran interés por el carácter placentero o doloroso de sus experiencias, y que esta preocupación se extiende hacia el futuro. Pero esta preocupación tiene una excepción, debida a que nuestro agente muestra una "Indiferencia hacia los Martes futuros"

"Durante todos los martes, se preocupa del modo habitual de lo que le ocurre. Pero nunca se preocupa por los posibles placeres o do-

³⁸ Ver capítulo III.

³⁹ Parfit (86), p. 121.

⁴⁰ Ver capítulo II, pp. 24 y ss.

lores de los martes futuros. De este modo, elegiría sufrir una operación muy dolorosa el siguiente martes en vez de una menos dolorosa el miércoles siguiente. Esta elección no sería el resultado de ninguna creencia falsa. Este hombre sabe que la operación será mucho más dolorosa si ocurre el martes. Tampoco tiene creencias falsas sobre la identidad personal. Esta de acuerdo en que será precisamente él quien sufrirá el martes. Tampoco tiene creencias falsas acerca del tiempo. Sabe que el martes es meramente una parte de un calendario convencional, con un nombre arbitrario tomado de una religión falsa. Tampoco tiene ninguna otra creencia que pudiera ayudarle a justificar su indiferencia hacia el dolor en los martes futuros. Su indiferencia es un hecho desnudo. Cuando hace planes para su futuro, simplemente le sucede que siempre prefiere la expectativa de un gran sufrimiento en martes a un dolor mediano cualquier otro día.⁴¹

Las preferencias de este hombre en lo que respecta a lo que le sucederá en los martes futuros es irracional. Y lo es porque prefiere lo peor sin tener ningún motivo para ello. Si tuviera alguna de las creencias mencionadas, sus preferencias seguirían siendo irracionales, pero el motivo sería la falsedad de sus creencias. Pero puesto que no las tiene su preferencia es intrínsecamente irracional, sencillamente porque lo peor suceda en martes no es ninguna razón para preferirlo y él sabe que no es ninguna razón. Por ello, y en este sentido, podemos decir que está racionalmente requerida la preocupación por los placeres y dolores de los martes. El motivo es que él se preocupa por sus placeres y dolores. Si no tuviera nunca esta preocupación, y pudiera justificar su actitud de tal modo que esto no supusiera que prefiere lo peor, por ejemplo alegando que una preocupación de este tipo debilita el carácter y esto es para él mucho peor que el dolor más grande, no habría esta exigencia. Pero la hay *porque él se preocupa*, porque para él el dolor es lo peor.

Algo parecido puede ocurrir en el caso de la moral. Supongamos que alguien tiene lo que Parfit llama "Altruismo de una milla de alcance". Esta persona siente un gran deseo de dar peso en sus consideraciones a los intereses de todos los que viven en un radio de una milla alrededor de su casa, pero no está dispuesto a darles ningún peso a los intereses de los que viven más allá. Y supongamos que este deseo no se basa en ninguna falsa creencia y que es puramente altruista, e.d., que no tiene ninguna base en el interés propio, por ello, en la conveniencia de coo-

⁴¹ Parfit (196), p. 124.

perar con los individuos próximos ni en la expectativa de sanciones de ningún tipo. Este preferencia es, en este sentido, una auténtica preferencia moral. Al igual que sucedía en el caso anterior, esta preferencia moral es irracional. Y por ello, en este sentido, está racionalmente exigido que dé peso a los intereses de todos, con indiferencia de cuán lejos habiten.

Pero lo importante es ver por qué esto está racionalmente exigido. Lo que está racionalmente exigido no es desear dar peso a los intereses de los demás, sino dar peso a los intereses de todos, supuesto que desear dar peso a los interés de algunos, sencillamente porque la cercanía espacial, al igual que la cercanía temporal, no es ninguna razón para preocuparse más. Es decir, lo que está racionalmente exigido es que las preferencias morales cumplan determinados requisitos, pero no el tener tales preferencias. Pero esto es justo lo que afirma T, a saber, que es cierto que hay un modo racional de elegir moralmente, pero no hay ninguna exigencia racional de adoptar el punto de vista moral. Y Parfit reconoce esto explícitamente pues afirma que "si alguien no tiene ninguna preocupación por lo que les sucede a los demás, esto, aunque deplorable, no puede ser irracional"⁴². Pero entonces ¿cómo se explica que Parfit deje abierta la cuestión acerca de si están o no exigidas racionalmente las preocupaciones de carácter moral? La respuesta es, sin duda, que Parfit, en este ejemplo, está hablando del *deseo*. El deseo de actuar moralmente no está racionalmente exigido. Lo que sí lo está es actuar moralmente⁴³.

Por tanto, la cuestión no es si CP afirma que está racionalmente exigido actuar moralmente, sino si hay alguna razón por la que debemos aceptar esta afirmación. Pero no parece posible dar una razón de este tipo⁴⁴. Y, aunque puede afirmarse que hay ciertos deseos que están exi-

⁴² Parfit (B6), p.125.

⁴³ Recuérdese que no estamos hablando aquí del interés personal en los asuntos ajenos sino del interés moral. El término 'deseo' puede resultar confuso en este contexto, si no recordamos que no estamos hablando de deseos personales, sino de la fuerza relativa de las distintas preferencias, ya sean personales o morales.

⁴⁴ Recuérdese que lo que sostiene T en este sentido es al que no es irracional actuar moralmente cuando nuestro mayor deseo es hacerlo así, e.d., si el mayor deseo de un agente es satisfacer un determinado criterio de conducta moral, entonces es racional que actúe para satisfacer tal criterio, supuesto que este criterio es racional, y por tanto, para tal agente, es racional aceptar los imperativos mora-

gidos racionalmente, esto sólo es cierto en el sentido de que un conjunto de deseos, por el hecho de contener determinados elementos, debe, so pena de irracionalidad contener otros, de aquí sólo puede deducirse que hay un modo racional de elegir moralmente, pero no que sea racionalmente exigible elegir desde el punto de vista moral⁴³. Si tengo determinadas preferencias morales, entonces debo tener otras. Pero no tengo por qué tenerlas en absoluto. Afirmar lo contrario es confundir la racionalidad de las preferencias con la exigencia racional de tener tales preferencias.

La afirmación (CP2) realizada por Parfit es la mantenida por Nagel⁴⁴, del cual Parfit toma el ejemplo⁴⁵. Para Nagel, lo que me da una razón para, por ejemplo, ayudar a alguien, no es el hecho de que yo sienta una especial simpatía hacia esa persona, o que me afecte su desgracia de algún modo personal, sino precisamente el hecho de que ese alguien necesite mi ayuda. Por tanto, puede mantenerse que yo tengo una razón para actuar moralmente, e.d., estoy racionalmente requerido a actuar moralmente, incluso si no tengo ningún deseo de hacerlo en absoluto.

La confusión a la que aludimos es la que da a estas afirmaciones un aire plausible. Tales afirmaciones son plausibles porque, en efecto, el hecho de que alguien necesite mi ayuda es parte de la razón para ayu-

tes; y b) que si un agente no quiere actuar moralmente, entonces para él no resultan obligatorios los imperativos morales, pues nadie puede ser acusado de irracional por no satisfacer un criterio que no quiere satisfacer. Para un estudio más detenido de este problema, ver Apéndice 2.0.

⁴³ En la tercera parte de *Reasons and Persons*, Parfit intenta otro camino para mostrar que la preocupación por los intereses de los demás está racionalmente exigida precisamente porque el conjunto de nuestros deseos incluye un interés por nosotros mismos. Básicamente, su argumento se apoya en una teoría poco habitual acerca de la identidad personal, según la cual no existe ningún sustrato permanente en la personalidad, sino que lo que llamamos "identidad personal" sólo consiste en la existencia de determinadas relaciones psicológicas. Esta visión humeana del yo como un mero sistema de fenómenos coherentes, si es defendible, puede suponer una objeción fundamental a todas las versiones de S, pues de ella se desprende que el hecho de que una experiencia vaya a ser mía en el futuro no es una razón para darle mayor importancia, igual que tampoco lo es el hecho de que determinada experiencia suceda en el futuro. Esta argumentación de Parfit es extremadamente interesante y puede resultar tremendamente fructífera. Sin embargo, no vamos a ocuparnos de ella aquí. El motivo es que aun queda otro camino, del que nos ocuparemos en el siguiente capítulo, para intentar mostrar por qué un agente racional debe ocuparse de la moralidad, sin necesidad de acudir a una teoría sobre la identidad personal tan poco habitual. Sólo si este otro camino más accesible se muestra infructuoso podrá plantearse como "último recurso" un análisis de los supuestos que subyacen a nuestra visión habitual sobre la identidad personal.

⁴⁴ Nagel (70).

⁴⁵ Parfit (1984), p.121.

darle. Pero no puede ser toda la razón. Que alguien necesite mi ayuda es una razón para ayudarlo siempre y cuando yo quiera satisfacer un cierto criterio racional de conducta moral. Por eso, si yo tengo ese deseo, la razón para ayudar a alguien es que ese alguien necesite mi ayuda. Pero ese hecho no es una razón en absoluto para desear ayudar si yo simplemente no tengo ningún deseo de actuar moralmente. Es decir, para que mis preferencias morales sean racionales, es necesario que yo prefiera ayudar a quien necesite mi ayuda que no hacerlo. Hay en este sentido una exigencia racional de tener estas preferencias morales y no otras. Pero no hay ninguna exigencia racional de tener preferencias morales en absoluto, en el sentido de desear satisfacer tales preferencias.

Podemos entonces concluir este capítulo diciendo que el intento de encontrar una teoría alternativa a T que no sea directamente auto-frustrante a nivel colectivo ha resultado infructuoso. En adelante, trataremos de analizar, partiendo de T, los motivos por los que un agente racional debe ocuparse de la moralidad.

CAPITULO XII

LA MORAL COMO RACIONALIDAD INDIRECTA

En el capítulo anterior vimos como no es posible encontrar una teoría alternativa sobre la racionalidad que no sea directamente auto-frustrante a nivel colectivo, debido a que para ello la tal teoría debe mostrar que esta racionalmente exigido actuar según preferencias morales, lo cual no parece fácil. Por ello, a lo largo de este capítulo intentaremos, partiendo de T, contestar a la pregunta acerca de los motivos que un agente racional, que no desea actuar moralmente, pueda tener para adoptar el punto de vista moral¹.

Si entendemos esta pregunta de modo general, entonces ya ha quedado contestada en términos negativos. En efecto, nuestra pregunta puede entenderse del siguiente modo. Hay situaciones en las que los dictados del interés y los de la moralidad están en conflicto. Podemos preguntar en estos casos que es lo que debe hacer un agente racional². En general, en estos casos un agente racional actuará según los dictados del interés propio. Por tanto, tenemos que matizar nuestra pregunta. Lo que queremos saber es qué debe hacer un agente racional cuando a) los dictados de la moral y del interés propio están en conflicto y b) si todos seguimos los dictados del interés propio el resultado será peor, en términos del interés propio, que si todos siguiéramos los dictados de la moral. Es decir, nuestra pregunta se limita a aquellas situaciones de dilema en las cuales nuestra teoría es directamente auto-frustrante a nivel colectivo.

También parte de esta pregunta restringida ha sido ya contestada negativamente, al decir que, en tales situaciones, la estrategia B es estrictamente dominante. Un agente individual no tiene ninguna razón en ta-

¹ Esto, como ya sabemos, no quiere decir que un agente racional nunca desee, por definición, actuar moralmente. Esto supondría considerar la conducta moral como irracional, visión que en el capítulo anterior hemos desechado como arbitrariamente restrictiva. Más bien, hemos admitido que, cuando el mayor deseo de un agente es satisfacer con su conducta un determinado patrón moral, entonces tiene una buena razón para actuar moralmente. Lo que preguntamos es qué motivos pueden dárseles a los agentes racionales que no tienen este deseo para actuar moralmente.

² En adelante y salvo que se diga expresamente lo contrario, al hablar de "agente racional" en este contexto nos referiremos a un agente racional que no desea actuar moralmente.

les casos para actuar moralmente. Si lo hiciera el resultado sería peor para él. Lo que esto quiere decir es que los dilemas no tienen solución en el nivel individual, cosa que no es en absoluto sorprendente, desde el momento en que el problema surge en el nivel colectivo. Este punto tiene una gran importancia. En primer lugar, esto es lo que significa decir que T es una teoría absolutamente satisfactoria a nivel individual, e.d., que es una buena teoría acerca de la racionalidad individual. En segundo lugar, tiene una gran importancia no sólo práctica sino también teórica, pues esta afirmación se deriva de la afirmación más general respecto a la importancia de la conducta ajena a la hora de determinar cuál es la conducta racional en una situación dada. En ocasiones se ha sostenido que cuál sea la conducta racional es un asunto "privado" que no depende de los demás y que puede determinarse de un modo individual e independiente. Imaginemos, por ejemplo, un atasco de tráfico producido por los propios conductores al detenerse en los cruces. Cada conductor individual razona del modo siguiente. Si los demás respetan los cruces, entonces yo llegaré antes si no los respeto y si los demás no los respetan, entonces también gano algo (por poco que sea) si yo tampoco los respeto³. Algunas veces se ha sostenido que, puesto que somos nosotros mismos los que producimos el atasco, lo racional es no contribuir a ello y respetar los cruces. A lo largo de todas las páginas anteriores hemos intentado demostrar precisamente lo contrario. Lo que es racional en cada momento depende básicamente de las circunstancias y, en las situaciones de interacción, la conducta de los demás es una parte esencial de las circunstancias.

Por ello, lo que deseamos saber es qué motivos puede tener un conjunto de agentes racionales para actuar moralmente. Esto no supone que haya una entidad abstracta llamada "comunidad" o "conjunto de agentes" con poder de decisión. En un sentido, todas las decisiones son tomadas por individuos. Lo que queremos saber es si, puesto que la conducta moral en los casos de dilema es colectivamente racional, los agentes individuales tienen algún motivo para comportarse moralmente y bajo qué supuestos

³ Dejamos de lado la cuestión de las situaciones con 'límites' en las cuales hay un individuo cuya acción, unida a acciones anteriores, es la determinante para producir el efecto indeseado. En primer lugar, en la práctica estos límites son difíciles, si no imposibles de señalar (¿cuántas personas tienen que pisar el césped para estropearlo?) y, en segundo lugar, este planteamiento introduce complicaciones innecesarias en este momento.

tienen estos motivos. Para ello, recurriremos a la noción de racionalidad indirecta, intentando ver el papel que esta puede desempeñar en la solución a nuestro problema.

Volvamos por un momento a nuestro ejemplo del taxista del capítulo X. Las alternativas a las que se enfrentaba nuestro agente eran, por orden de más a menos preferida, las siguientes:

1. El taxista me lleva y no pago.
2. El taxista me lleva y pago.
3. El taxista no me lleva y no pago.

Debido a las características de la situación, el resultado preferido 1 no era en absoluto una alternativa. De modo que lo mejor que nuestro agente podía hacer era intentar conseguir 2. El problema era que, precisamente por ser un agente racional, 2 no podía lograrse, a menos que el agente hiciera algo. Vimos que a este respecto podían hacerse dos cosas. Podía transformarse la situación de tal modo que la solución 2 pasara a ser la alternativa preferida a nivel personal. A esto le llamábamos "solución R". Pero también podía operar una transformación en sí mismo, de tal modo que su mayor deseo fuera pagar sin considerar si hacerlo le resulta o no beneficioso. A este le llamábamos "solución I"⁴. Ambas soluciones son ejemplos del método indirecto de atarse uno mismo.

Veamos ahora lo que ocurre en un caso de dilema, por ejemplo en el dilema bipersonal del prisionero. Las alternativas a las que cada agente se enfrenta, por orden de preferencia, son

1. Tu no confiesas y yo confieso (A_2, B_1)
2. Ninguno de los dos confesamos (A_1, B_1)
3. Los dos confesamos (A_2, B_2)
4. Yo no confieso y tu sí confiesas (A_1, B_2)

donde yo soy el jugador A, tú el jugador B, A_1 y B_1 nuestras respectivas estrategias de no confesar y A_2 y B_2 las de confesar. Al igual que ocurriría en el caso del taxista, las circunstancias hacen que el resultado 1 sea inalcanzable. En este caso, la circunstancia culpable de esto es el

⁴ Tras lo dicho en el capítulo anterior, es evidente que no debe entenderse que la actuación según preferencias morales sea necesariamente irracional. Una vez que yo he adquirido preferencias morales, si actuación según estas no es irracional. De modo que en un sentido la solución 1 al caso del taxista es similar a la solución R del caso del ladrón, pues ambas consisten simplemente en cambiar el sistema de las preferencias sin hacer que estas se transformen en irracionales. Sin embargo, podemos seguir utilizando el término "solución I" para referirnos a la solución consistente en la adquisición del deseo de actuar moralmente pues a) no está racionalmente exigido el actuar según preferencias morales y b) desde el punto de vista del interés propio, e.d., de las preferencias personales, cuya satisfacción constituye el auténtico deseo del agente, la actuación según preferencias morales es irracional.

hecho de que la estrategia "confesar" sea dominante. Este hecho también hace inalcanzable el resultado 2. De modo que parece que sólo podremos conseguir 3. Pero podemos hacer algo mejor. Podemos intentar conseguir 2. Podemos intentar llegar a un acuerdo cooperativo. Sin embargo, no puedo intentar conseguir 1. Sencillamente, porque tú no vas a dejarte. De modo que, si bien 1 no es una alternativa en absoluto, 2 sí lo es. Al igual que ocurría en el caso del taxista, también ahora podemos aplicar soluciones del tipo R, que serían lo que nosotros hemos llamado soluciones políticas, consistentes en introducir mecanismos destinados a hacer im-ponibles los acuerdos, e.d., a introducir cambios en la situación que cambien la utilidad asociada con cada una de las estrategias disponibles. Esta es, sin duda, una forma de atarse uno mismo y, al mismo tiempo, atar a los demás, tal y como requiere la solución a un problema colectivo. Pero, como vimos en su momento, este tipo de soluciones no siempre son posibles y a menudo son muy costosas. Las soluciones consistentes en que todos los agentes actúen moralmente son mucho más eficaces y poco costosas. Supongamos que los agentes ordenan sus preferencias de un modo distinto. Por ejemplo supongamos que cada uno establece este orden

1. Ninguno de los dos confiesa (A_1, B_1)
2. yo confieso, tú no (A_2, B_1)
3. los dos confesamos (A_2, B_2)
4. yo no confieso, tú sí (A_1, B_2)

Es decir, cada uno prefiere confesar si piensa que el otro también lo hará, pero prefiere no confesar si piensa que el otro tampoco lo hará. Una situación en la que los agentes muestran estas preferencias ya no es un dilema, sino un juego distinto, al que Sen llama *Assurance Game*⁵. Este juego tiene dos puntos de equilibrio, (A_1, B_1) y (A_2, B_2), y cuál de ellos sea el resultado depende de lo que cada jugador espere de la conducta del otro. Llamemos a estas preferencias "preferencias AG", y a las preferencias en la situación de dilema "preferencias PD". Supongamos que los agentes actúan como si tuvieran preferencias AG. Entonces sería posible alcanzar un resultado óptimo en términos de las preferencias AG. Y, lo que es más importante, de alcanzar este resultado la situación también sería óptima en términos de las preferencias PD, e.d., de las preferen-

⁵ Sen (74), p.59.

cias reales de los agentes. Pero todavía es posible algo mejor. Es posible obtener un resultado óptimo con seguridad. Supongamos una ordenación distinta de las preferencias, según la cual cada agente prefiere

1. Ninguno de los dos confesamos (A_1, B_1)
2. Yo no confieso y tú sí (A_1, B_2)
3. Yo confieso y tú no (A_2, B_1)
4. Los dos confesamos (A_2, B_2)

Llamemos a estas preferencias "preferencias OR"⁴. En este nuevo juego A^1 y B_1 son estrategias estrictamente dominantes. De modo que, si actuamos como si tuviéramos estas preferencias, obtendremos con toda seguridad un resultado que no sólo será óptimo en términos de las preferencias OR, sino también de las preferencias DP, e.d., nuestras auténticas preferencias.

En un sentido, esto no es nuevo, pues significa lo que ya sabemos, es decir, que si tuviéramos preferencias morales entonces habría una solución para situaciones que, de otro modo, serían dilemas. Lo nuevo de este planteamiento es lo siguiente. Aun en el caso de que no tengamos preferencias morales, sabemos que, de tenerlas o de ser capaces de actuar como si las tuviéramos, entonces el resultado sería mejor no sólo para las satisfacción de las preferencias morales, sino de nuestras preferencias personales. Es decir, planteado de esta manera resulta fácil ver que esta situación tiene todas las características para permitir la aplicación de los mecanismos de racionalidad indirecta, pues

- a) lo mejor es tomar en el momento t_1 una decisión, la de cooperar, y llevarla a cabo en el momento t_2 , y, sin embargo
- b) cuando llega el momento de llevar a cabo la decisión los agentes no son capaces de hacerlo.

Y esto se mantiene para los casos en los que las soluciones indirectas R no son aplicables o son prohibitivamente costosas. En el artículo mencionado, Sen plantea la cuestión de una forma similar. Tenemos tres ordenaciones posibles de las preferencias. Ahora bien, Sen supone que la moralidad, al menos en tanto que esta tiene que ver con la consecución de una situación social óptima, ordenaría estas distintas ordenaciones alternativas de tal modo que la ordenación OR sería la preferida en primer

⁴ Sen (74), p.60.

lugar, seguida de la ordenación AG y, por último de la ordenación PD. Es decir, Sen propone considerar la moralidad como un mecanismo de ordenación, no de alternativas de acción, sino de ordenaciones de estas alternativas, e.d., como una cuasi-ordenación de ordenaciones u ordenación de segundo grado⁷. Una vez que consideramos así la moralidad, el problema resulta paralelo a los casos analizados de debilidad de la voluntad, pues tenemos una ordenación que es, desde el punto de vista moral, la que deseamos tener, pero que no es sin embargo la que realmente tenemos.

Este modo de ver el problema es interesante, pero no parece que resuelva nuestro problema, aunque sí lo clarifica. En efecto, desde el punto de vista moral, pueden ordenarse las ordenaciones. Es decir, si tenemos dos conjuntos de ordenaciones (tres en el caso presentado por Sen), tal que uno corresponde a la ordenación realizada desde el punto de vista moral, desde luego que desde este punto de vista esta ordenación es la preferible. La novedad aquí consistiría en que en este modelo no se considera que la moralidad sea una cuestión de todo o nada, sino que se permite una gradación de las ordenaciones alternativas de más a menos morales, o de mejores o peores desde el punto de vista moral, cosa que para Sen expresa mejor nuestra ideas intuitivas acerca de la moralidad. Pero, sea como sea, para nuestro problema esto no presenta especiales ventajas, pues nuestro propio modelo, al permitir también la existencia de ordenaciones distintas, a saber una ordenación desde el punto de vista moral y una ordenación desde el punto de vista personal, también permite la posibilidad de contemplar el problema como una cuestión de debilidad de la voluntad. Veamos como puede plantearse el problema en estos términos.

Supongamos que un agente quiere satisfacer un determinado requisito moral mediante su conducta, y que su mayor deseo consiste en satisfacer tal requisito. T afirma que lo racional para este agente es actuar de tal modo que sus preferencias morales se vean satisfechas. Es decir, lo racional es tomar la decisión de actuar según estas preferencias y llevar a cabo esta decisión. Sin embargo, él sabe que cuando llega el momento no puede hacerlo. Es incapaz de desoir la voz de sus preferencias

⁷ Sen habla de "cuasi-ordenación" porque entiende que la ordenación desde el punto de vista moral no tiene por que ser completa. Esto es muy discutible, al menos si admitimos, como parece que debemos hacer, la posibilidad de la indiferencia moral. En cualquier caso, no pretendo discutir esto aquí pues es indiferente para nuestro propósito actual.

personales. Sus preferencias personales pueden ser tan racionales como se quiera y, en este sentido, no sería irracional actuar según ellas. Lo que hace irracional esta actuación es que el deseo predominante del agente es satisfacer, no sus preferencias personales, por muy racionales que sean, sino sus preferencias morales. En este sentido puede decirse que es irracional actuar según las preferencias personales y que hacerlo así sólo puede explicarse acudiendo a la debilidad de la voluntad.

Hasta cierto punto, el problema de las preferencias personales como "tentación" se plantea en todos los casos en los que el agente desea actuar moralmente. En muchos casos la tentación puede ser superada. Pero hay casos en los que esto no sucede. En estos, un agente racional puede acudir a las estratagemas de la racionalidad indirecta. Son casos en los que la voluntad del agente es tan débil que sólo es capaz de actuar según sus preferencias personales. Puede entonces intentar seguir uno de los dos caminos de la racionalidad indirecta. Por un lado, puede introducir cambios en la situación de tal modo que la estrategia aconsejada desde el punto de vista moral sea también la mejor personalmente. Por otro, puede introducir un cambio en sus propias preferencias, convirtiendo sus preferencias morales en preferencias personales. Habitualmente, el primer camino no es practicable y, cuando lo es, a menudo resulta excesivamente costoso, de modo que la mejor solución suele consistir en alterar las propias preferencias. Supongamos que es esto lo que de hecho se hace. Por ejemplo, supongamos que el agente se expone a sí mismo a determinadas influencias externas con el propósito de aumentar su capacidad de simpatía. Una vez que ha conseguido que la elección que mejor satisface sus preferencias morales sea al mismo tiempo la que mejor satisface sus preferencias personales, habrá eliminado el problema. Podrá obrar según sus preferencias personales en una situación en las que estas ya no están en conflicto con sus preferencias morales.

Sin embargo, es importante subrayar algo. Formalmente, su elección será una elección desde el punto de vista personal. Pero en realidad su elección será una elección moral, pues son sus preferencias morales las que se intenta satisfacer mediante la puesta en práctica de estratagemas

indirectas⁸. Paralelamente, actuar según sus preferencias personales es, en cierto sentido, irracional, puesto que su mayor deseo es actuar desde un punto de vista moral. Pero en un sentido más importante su acción es racional, aunque esta racionalidad sea indirecta.

Ahora bien, la posibilidad de plantear este problema en términos de debilidad de la voluntad sólo resulta útil para la resolución de nuestro problema de modo indirecto. Lo que Sen muestra es que la elección entre las distintas ordenaciones alternativas puede ser considerado como un problema de debilidad de la voluntad *desde el punto de vista moral*. Pero nuestro problema es precisamente el que se le plantea a un agente racional que no tiene interés en satisfacer ningún requisito moral y para el que, por consiguiente, lo racional es actuar para satisfacer del mejor modo posible sus preferencias personales. De modo que la sugerencia de Sen sólo es aplicable a nuestro problema si podemos darle la vuelta. Es decir, no se trata de mostrar que desde el punto de vista moral la ordenación de las preferencias según un criterio moral es superior y que un agente puede encontrarse en un caso en el que su ordenación real de preferencias no coincida con ella, sino de mostrar que, en un sentido, la ordenación moral de las preferencias también es deseable desde el punto de vista personal y que, por tanto, cuando el agente tiene una ordenación distinta el problema también puede plantearse en términos de debilidad de la voluntad.

Volvamos al dilema del prisionero. Cada uno de los agentes tiene unas determinadas preferencias personales, y si actúa según ellas el resultado será subóptimo. Sin embargo, si tuvieran preferencias morales y actuaran según ellas, entonces conseguirían un resultado que no sólo sería óptimo con respecto a esas preferencias morales, sino también con respecto a sus respectivas preferencias personales. Podemos decir por tanto que desde el punto de vista personal es preferible actuar como si se tuvieran preferencias morales. Es decir, cada uno de los agentes tiene

⁸ En el caso presente de las preferencias morales, esto no tiene mucha importancia, al menos si suponemos que el objetivo de la moral, y por ende lo que hace que una acción sea deseable desde el punto de vista moral, es primordialmente la consecución de determinados resultados, de modo que no tiene mucho sentido preguntar si la elección de la estrategia moral se debe a que esta era también la preferible desde el punto de vista personal o no. Sin embargo, conviene señalar esto por la importancia que el problema tendrá posteriormente al tratar las preferencias personales.

un motivo para actuar moralmente suponiendo que todos los demás se comporten también moralmente.

Sin embargo, actuar moralmente en una situación de dilema es irracional. Lo racional es tomar la decisión de cooperar. Pero al llegar el momento de llevarla a cabo, lo racional es no hacerlo. Esto es lo que sucedía en el caso del taxista, en el que no es la debilidad lo que me impide llevar a cabo mis decisiones, sino mi propio carácter racional. En tanto que el problema es el mismo, es susceptible del mismo tipo de soluciones, a saber, las catalogadas como estratagemas indirectas. Puesto que soy racional, sólo podré llevar a cabo mi decisión de actuar moralmente si cambio mis preferencias. Por lo tanto, lo racional es acudir a los métodos disponibles para efectuar este cambio. Ahora bien, lo que distingue este caso del caso del taxista, y hace de él un problema de racionalidad colectiva y no de racionalidad individual es que el problema no se soluciona si yo cambio mis preferencias sino si todos lo hacemos. Naturalmente, si los demás cambian sus preferencias y se comportan de una manera moral, entonces lo racional para mí es no cooperar. Pero el problema, y al mismo tiempo la solución, está en que los demás sólo accederán a cambiar sus preferencias si yo también lo hago.

Tenemos pues que la ordenación de las alternativas según la preferencia moral es preferible no sólo desde el punto de vista moral sino también desde el punto de vista personal. Si bien nadie tiene motivos para querer cambiar sus preferencias en esta dirección, todos tienen un motivo para hacerlo supuesto que esto es condición para que los demás también lo hagan. Esto no es nuevo. Como ya sabemos, la elección de la estrategia cooperativa sólo es racional para un agente si su cooperación es condición indispensable para asegurar la cooperación de los demás. Nosotros sólo cooperaremos si estamos seguros de que los demás lo harán, pero no porque en el supuesto de que los demás cooperen lo racional sea cooperar, sino porque los demás sólo cooperaran si están seguros de que yo lo voy a hacer. Supongamos que en el caso del prisionero disponemos de una droga capaz de hacernos cooperar. Ambos tenemos una razón para comprometernos a tomar la droga. Pongámonos en la situación de uno de ellos. Sólo si el otro toma la droga la tomaré yo también. Naturalmente, si el otro la toma yo haré mejor no tomándola. Pero es que el otro sólo la tomará si me la ve tomar a mí.

Precisamente es esta exigencia de reciprocidad lo que hace posible que todos tomemos la droga y solucionemos así el problema. Pero también limita el rango de las soluciones disponibles. Tal como queda planteada la situación, es evidente que las mejores soluciones son lo que llamamos soluciones políticas, precisamente porque estas son soluciones colectivas y que, por ser externas, son capaces de garantizar el cumplimiento de la cooperación. Y las soluciones políticas representan, como ya dijimos, el ejemplo paradigmático del método de atarse uno mismo en el nivel colectivo. Pero, como también sabemos ya, las soluciones políticas no siempre son las mejores. Sin embargo, esto hay que matizarlo. Lo que no es siempre lo mejor es alterar la situación. En muchas ocasiones es mejor alterar el orden de las preferencias. A esta alteración le llamábamos solución moral, pues consiste básicamente en que los agentes adquieran el deseo de actuar según preferencias morales. Pero en un sentido estas soluciones también deben ser soluciones políticas, a saber, en el sentido de que la decisión de producir esta alteración debe ser tomada colectivamente y su cumplimiento debe ser asegurado. Y esto es así porque la cooperación sólo es posible si transformamos la situación de tal modo que hagamos que la cooperación de uno sea condición indispensable para la cooperación del otro. En este sentido, Harsanyi tiene razón al afirmar que la cooperación sólo es racional, y por tanto la solución cooperativa sólo es alcanzable, si los acuerdos son imponibles.

En la introducción a la tercera parte de este trabajo vimos las distintas soluciones a las situaciones de dilema, agrupándolas según los motivos por los que un agente podía escoger la estrategia A (truista). Tanto las soluciones políticas como algunas de las soluciones internas tienen como consecuencia abolir el dilema, haciendo que la estrategia A sea también la mejor en términos de interés propio. Cuando se dan esas situaciones, el problema se resuelve porque la elección moral y la elección personal dejan de estar en conflicto. El problema teórico interesante surge en la solución 5, en la cual el problema práctico se resuelve porque los agentes actúan moralmente, e.d., escogen la estrategia A pesar de que esto es peor para ellos. Esto supone un problema teórico porque esta solución consiste en que los agentes se comporten de una manera irracional. Lo que hemos intentado mostrar es que este comportamiento por parte de los agentes es indirectamente racional. Comportarse moralmente

es indirectamente racional del mismo modo que lo era para nuestro agente del caso del taxista convertirse a sí mismo en un hombre de confianza. Y es tan indirectamente racional como lo es la introducción de medidas políticas. La diferencia está en que, mientras que tanto las soluciones políticas como las soluciones 3 y 4, consistentes en la transformación de las preferencias personales en preferencias morales, son casos paralelos a las llamadas "soluciones R" en los casos individuales, la solución 5, consistente en que todos actuemos de un modo moral, e.d., eligiendo A con independencia del resultado que esto tenga para nosotros, es paralela a las soluciones I. Lo característico de estas soluciones, como se recordará, es que el agente se ata a sí mismo con el fin de ser capaz de realizar actos irracionales*, con el fin de satisfacer del mejor modo posible sus objetivos racionales. Y es esta finalidad la que permite llamar racional a esta conducta, aunque su racionalidad pase por la actuación irracional.

Esto puede comprenderse con mayor facilidad si suponemos una comunidad de individuos racionales, sin ningún interés por la moralidad, que saben que van a verse enfrentados a dilemas. Saben que para conseguir un resultado óptimo tienen que poner en práctica alguna de las soluciones de nuestro cuadro. Supongamos para simplificar que las soluciones políticas 1 y 2 no son siempre aplicables o resultan demasiado costosas. Es mucho mejor intentar acceder a las soluciones 3, 4 y 5. Supondremos aquí que estas soluciones son igualmente fáciles (o difíciles) de introducir. Podemos suponer también que es mejor intentar aplicarlas todas ellas y no centrarse en una sola, debido a que los distintos individuos, a causa de sus distintas capacidades y disposiciones, muestran una facilidad mayor para una de estas que para las otras. Algunos individuos tienen más imaginación y una mayor capacidad de simpatía, por lo cual les resulta poco costoso ponerse en las situación de los demás y compadecerse de ellos hasta tal punto que, con gran facilidad, llegan a hacer suyos los intereses de los demás. A tales individuos, especialmente sensibles al dolor y el placer ajeno, resulta fácil educarles para conseguir la solución 3, en

* Debe recordarse siempre que estos actos son irracionales no por su propio carácter sino debido a que el agente no tiene interés en su realización. Sólo en este sentido puede decirse que tales actos sólo pueden ser indirectamente racionales.

la cual la elección de la estrategia A se realiza debido a que esta estrategia es también la mejor desde el punto de vista personal. Por otra parte, hay individuos con una gran disciplina y poca imaginación, con gran capacidad para los conceptos abstractos a los que es más fácil inculcarles el sentido del deber. Entre ellos, algunos son especialmente orgullosos y pagados de sí mismos, tanto que el sólo hecho de pensar que están obrando por sentido del deber o por algún otro motivo de esta índole es suficiente para dejarles tan satisfechos que, aunque están dispuestos a elegir la estrategia A sean cuales sean sus consecuencias, al actuar así la estrategia A se convierte automáticamente en la mejor para ellos desde el punto de vista personal. Para ellos la solución 4 es más accesible que las otras. Otros sin embargo, aunque también capaces de actuar por sentido del deber, no tienen tal concepto de sí mismos ni les preocupa tanto sentirse buenos y cooperadores. Para ellos la solución 5 es la ideal. Es decir, puede educárseles con relativa facilidad para actuar moralmente, e.d., para elegir A pesar de que positivamente saben que esto es peor para sí mismos.

Supongamos pues que los miembros de esta comunidad deciden poner en práctica todas las estratagemas indirectas posibles para llevar a cabo estas transformaciones en las preferencias de los individuos. Naturalmente, como son individuos racionales cuya meta es alcanzar un resultado óptimo, sólo se someterán a la influencia de estos mecanismos a condición de que lo hagan todos. Ahora bien, la cuestión fundamental es si tenemos algún motivo para aducir que los individuos que se transformarán del modo indicado por las soluciones 3 y 4 son más racionales que los que lo harán del modo indicado en 5. La respuesta es evidentemente negativa. Es cierto que los primeros harán elecciones racionales, mientras que los últimos las harán irracionales. Pero lo fundamental es que en todos los casos el objetivo es racional y sólo se prestarán a transformarse de ese modo debido a su interés en este objetivo. Esto es lo que sucedía en el caso del ladrón, en el cual teníamos la solución R, realizada mediante la droga2 y la solución I droga1. El individuo que tomaba la droga1 se transformaba en un individuo irracional, mientras que el que tomaba la droga2 simplemente adquiría unas preferencias distintas, pero racionales, y se comportaba racionalmente según la satisfacción de su nuevo conjunto de preferencias. Si suponemos que en la misma situación un individuo toma la dro-

ga2, mientras que el otro, alérgico a este fármaco, toma la droga1, veremos que no hay ningún motivo para suponer que uno de ellos es más racional que el otro. Lo fundamental es que ambos toman la droga porque son individuos racionales que quieren satisfacer del mejor modo posible sus preferencias racionales.

Podemos entonces concluir del modo siguiente. El comportamiento moral, la elección de la estrategia altruista, es individualmente irracional. Pero a nivel colectivo no lo es. Es decir, en una situación de dilema la conducta racional es transformar la situación de tal modo que se asegure la cooperación, haciendo de la cooperación de cada uno condición necesaria para la cooperación de los demás. En ocasiones, esto pasa por operar en los agentes la transformación consistente en adquirir un deseo incondicionado por actuar moralmente. Cuando esto sucede, el comportamiento moral es indirectamente racional para cada uno de los agentes individuales. Y el hecho de que esto sea así es lo que soluciona el problema teórico que supone la solución moral de los dilemas para la teoría de la racionalidad.

El objeto de esta tesis ha sido el estudio de las relaciones entre moralidad e interés. Puesto que las conclusiones de este estudio ya han sido resumidas al final del capítulo XII, quisiera ahora señalar brevemente las limitaciones del mismo.

En primer lugar, es preciso insistir en que el objeto del estudio han sido las relaciones conceptuales entre moralidad e interés. En ningún momento se ha intentado apuntar a los posibles mecanismos para poner en práctica la solución moral de los dilemas. La droga que hemos utilizado en nuestros ejemplos (ver p.273, por señalar un caso) no existe. Lo único que se ha pretendido mostrar es que, de existir, sería racional que todos la tomáramos.

En segundo lugar, es importante señalar algunos asuntos que, aunque pueden englobarse en el estudio de lo que hemos llamado relación conceptual, han quedado sin tratar. Parte del motivo por el que no nos hemos ocupado de ellos es su conexión con los problemas prácticos, pues su solución depende en gran medida de los medios con los que se cuente para operar en los agentes el cambio moral. Puesto que es improbable que podamos disponer alguna vez de una de nuestras drogas, es previsible que los mecanismos utilizables en la práctica tengan que ver con la educación. Y aunque ignoremos los medios específicos que pueden emplearse en esta educación, es previsible que surjan algunos problemas, al menos si estos medios procuran una educación semejante a la que ahora conocemos. Quisiera especialmente referirme a dos de ellos, que me parecen particularmente importantes.

El primero hace referencia a la duración del cambio moral. A diferencia de lo que sucedería en el caso de las drogas, los efectos de la educación moral, al menos tal y como la conocemos, no son pasajeros. Si una comunidad es educada para actuar moralmente, actuará de este modo en todas las ocasiones. En este punto cabe preguntarse si actuar moralmente es siempre racional. Naturalmente, lo es en todos los casos de dilema. Por tanto, si identificamos actuar moralmente con respetar los acuerdos cooperativos el problema no se presenta. Pero esto es sólo una parte de lo que normalmente se entiende por conducta moral. lo que no sólo plantea el problema de qué tipo de moral ha de enseñarse, sino también, y fundamentalmente, si es posible enseñar una moral que sólo consista en eso. La

regla moral que garantiza mínimamente el cumplimiento de la identificación mencionada y que, por ello, asegura que la conducta moral será siempre racional debe ser del tipo "Elige la estrategia A (altruista) sólo en aquellas situaciones de dilema en las cuales sea posible garantizar el cumplimiento de un acuerdo condicional". El problema por tanto no es ya sólo identificar situaciones de dilema, lo que ya es bastante difícil en la práctica, sino determinar si se dan el resto de las condiciones que permiten aplicar la regla.

Sin embargo, tales problemas no sólo se plantan respecto a la conducta moral sino también respecto a la conducta racional. Por ejemplo, la misma dificultad aparece cuando se trata de saber si uno va a enfrentarse con un dilema iterativo o no. Por tanto, y puesto que además la conducta moral no es sino un método de racionalidad indirecta (al menos si aceptamos la identificación anterior) podemos suponer que deben emplearse los mismos métodos de cálculo de probabilidades y asignación de utilidades para saber cuándo debe elegirse la estrategia altruista. Esto no pasa de ser un apunte general, pero puede ser una base de partida para un estudio profundo del problema.

El otro tema que queda sin resolver es si puede obrarse moralmente a sabiendas de que esto es únicamente una estratagema indirecta. La respuesta depende en gran medida de la solución al problema anterior. Desde luego, si es posible inculcar una regla moral que respete la identificación anterior y cuyo cumplimiento sea siempre indirectamente racional, entonces la respuesta es afirmativa. El problema surge si la regla moral que se enseña no es del tipo señalado más arriba, sino del tipo "Elige siempre la estrategia A sin tener en cuenta los posibles resultados que esto tenga para ti". Es perfectamente posible que la educación moral y, por tanto, la garantía del cumplimiento de los acuerdos cooperativos sólo puedan realizarse apoyándose en una regla de este tipo, que tiene el problema de ocultar el carácter de estratagema de la conducta moral. Si esto sucediera, pudiera resultar que el único modo de obrar racionalmente, e.d., de conseguir los objetivos propuestos por nuestra teoría de la racionalidad, sería no sólo utilizar estratagemas indirectas sino, además, hacerlo sin saber que se hace. De ser así, esto platearía un problema adicional a la teoría, cuya posible solución requiere un estudio por sí mismo.

En último lugar, quisiera quisiera hacer hincapié en las limitaciones de la presente tesis. Concretamente, quisiera aclarar lo que no se pretende con ella. Por un lado, no se pretende de ningún modo que este sea el único fundamento posible de la moralidad. Ni siquiera se pretende que sea el único fundamento racional, e.d., que sea imposible mostrar que la conducta moral está racionalmente exigida. Lo único que he mantenido es que, partiendo de una teoría concreta sobre la racionalidad práctica (a la que hemos llamado T) que parece buena y defendible por los motivos analizados, no parece que esto sea así.

Por otro lado, no se ha intentado dar una definición satisfactoria de lo que es moral o de en qué consiste la conducta moral. Naturalmente este es un tema de gran importancia, sin duda mayor que la del asunto del que en esta tesis nos hemos ocupado. Pero dado lo restringido de nuestro objetivo, tal y como ha quedado expresado al principio de estas conclusiones, nos hemos limitado a una caracterización básica de la conducta moral como aquella que es imparcial entre los intereses de todos, caracterización que, si bien es a todas luces incompleta, creo que es básicamente adecuada y tiene la ventaja de poder ser aceptada como punto de partida sin demasiados problemas.

El objeto de exponer estas limitaciones no es justificar las carencias del trabajo, ni se espera con ello benevolencia, pues, si bien es cierto que hay que valorar las cosas por lo que son, es igualmente cierto que también debe valorarse lo que pudieron ser y no son. Mas bien lo que espero es que tener en cuenta estas limitaciones ayude a comprender el alcance de la presente tesis y la importancia del tema tratado, que no es sino la aludida relación conceptual entre la conducta moral y una determinada teoría acerca de la racionalidad práctica.

APENDICES

APENDICE 1.

Aunque es indudable la existencia de una multiplicidad de fines intermedios, cuando nos trasladamos al nivel de los fines últimos, es decir, aquellos que son queridos en virtud de sí mismos y no de otra cosa, la cuestión acerca de si hay un sólo fin último o varios se vuelve problemática. Hay multitud de cosas que, aparentemente, son queridas por sí mismas, tales como ver determinada película, tomar un baño de sales o leer un libro. Sin embargo, se ha mantenido repetidamente que esta aparente multiplicidad puede ser reducida a un único fin último.

La reducción puede plantearse de dos formas. En primer lugar, puede hablarse de una reducción en sentido aristotélico, en la cual se entiende que los diversos objetivos que los hombres se plantean, si bien son fines particulares de una acción determinada, no son fines últimos, pues no son queridos por virtud de sí mismos, sino por alguna otra cosa con respecto a la cual son a su vez medios. El razonamiento continúa señalando que es imposible que esta cadena se extienda hasta el infinito, por lo cual es necesario que exista un fin último que nunca sea buscado por otra cosa, sino siempre por sí mismo, y a partir del cual adquieran valor los fines intermedios. Por lo tanto, puede afirmarse que, si bien es cierto que en el nivel de los fines intermedios o subordinados existe una multiplicidad, esta desaparece en el nivel del fin último.

Sin embargo, esta conclusión no parece seguirse necesariamente del razonamiento expuesto. Es cierto que algo que es fin particular de una acción puede no tener valor por sí mismo, sino sólo porque es a su vez medio para alguna otra cosa. Así por ejemplo, si bien el fin inmediato de mi acción de salir de casa y dirigirme a la parada del autobús es tomar este, este fin no tiene valor en sí mismo, sino sólo en tanto que es medio para el fin ulterior de ir a la Puerta del Sol, y a su vez, esto sólo tiene valor en tanto que medio para el fin de ir al cine y ver una determinada película. También podemos dar por cierto que esta cadena debe parar en algún sitio. Pero esto no significa que sólo deba existir un fin último.

Si es cierto que hay una multiplicidad de fines intermedios, también parece que hay una multiplicidad de fines últimos, es decir, de objetivos queridos por sí mismos y no en virtud de alguna otra cosa. Por ejemplo,

ver una determinada película puede ser un fin último en este sentido, como también puede serlo en algún momento pasear por la Puerta del Sol, leer un libro o bailar. La conclusión aparente de que la multiplicidad puede ser reducida a un único fin último surge a partir de una reflexión ajena al razonamiento expuesto, aunque relacionada con él.

Todos los fines que se puedan proponer *pueden* ser fines intermedios, como es evidente en el caso aludido del fin de ir a la Puerta del Sol, que si bien puede ser querido por sí mismo, también puede ser querido en virtud de otra cosa. Al intentar buscar el fin último cuya existencia es necesaria según el razonamiento que hemos aceptado como válido, puede pensarse que lo que se necesita para que algo sea un fin último no es sólo algo que no sea querido por otra cosa sino que *nunca* adquiera su valor en virtud de otra cosa.

Sin embargo, esto es un supuesto adicional que no está exigido por el razonamiento acerca de la necesidad de algo que sea un fin último y que, por tanto, no puede ser defendido apelando a este. Lo único que el razonamiento demuestra es que no todo puede ser fin intermedio, sino que es necesario que algo sea un fin último, pero de ninguna manera se sigue que el fin último haya de ser único.

En *segundo* lugar, puede plantearse una reducción que puede constituir un segundo paso de la reducción anterior pero que también puede surgir de un modo independiente, y a la cual podemos referirnos con el nombre de "reducción substantiva" para señalar que su propósito es aislar aquello que constituye propiamente el fin último. El argumento anterior, unido al supuesto de que sólo aquello que nunca es querido por otra cosa puede ser fin último, reduce enormemente la lista de posibles candidatos a ocupar el puesto de fin último. Pero, aun así, parece que en principio es posible que existan varios fines que cumplan este requisito de no poder ser queridos mas que por sí mismos. De hecho, el propio Aristóteles, tras exponer el razonamiento que concluye con la necesidad de la existencia de un fin último, sigue dejando abierta la posibilidad de que este no sea único¹. Sin embargo, inmediatamente después, Aristóteles identifica la felicidad como el único fin último de nuestras acciones. Esta identificación no parece seguirse de ningún razonamiento teórico,

¹ Aristóteles, *Ética a Nicómaco*, 1097a

sino que más bien parece basarse en un hecho empírico: la felicidad cumple el requisito de no ser nunca elegida por otra cosa, sino siempre en virtud de ella misma y es además lo único que cumple esta condición, pues todo lo demás se elige siempre por ella, siendo por tanto la felicidad fin de todos los actos. Tenemos por tanto una reducción completa a un único fin último. Podíamos rechazar sin más la existencia de un único fin último simplemente rechazando como arbitrario el supuesto de que un fin último ha de ser algo que nunca es querido por otra cosa. Sin embargo, es importante ver que, aun dando por bueno ese supuesto, puede rechazarse esta segunda reducción a un único fin último, llámese este felicidad o cualquier otra cosa².

La afirmación de que es la felicidad el único fin último pretende basarse en un hecho empírico. Sin embargo, a nivel empírico es fácil comprobar que hay algunas otras cosas que parecen ser también siempre deseadas por ellas mismas. El propio Aristóteles menciona algunas de ellas, por ejemplo los honores y el placer. Sin duda, esto no es verdad sin matices. Es cierto que algunas personas nunca buscan los honores por otra cosa sino siempre por ellos mismos, pero aunque a nivel intrapersonal los honores o el placer puedan cumplir este requisito no puede decirse lo mismo a nivel interpersonal, pues es indudable que hay personas que nunca, o al menos no siempre, buscan los honores por sí mismos, sino que lo hacen por alguna otra cosa, como por ejemplo por el poder o los privilegios sociales que estos traen consigo. Vemos pues que el supuesto de que el fin último ha de ser siempre buscado por sí mismo puede presentarse en dos versiones, una a la cual llamaremos versión intrapersonal y otra a la cual nos referiremos como versión interpersonal. Podemos preguntar cual de las dos versiones estamos utilizando. Sin embargo, no es necesario buscar una respuesta a esta pregunta, pues aun en el caso de que entendamos el supuesto en su versión débil el argumento de la reducción substantiva mantiene su afirmación de que sólo existe un único fin último. Y esto a pesar de que se reconoce que hay otras cosas que, al menos a nivel intrapersonal, son siempre buscadas por sí mismo. El argumento que res-

² El argumento siguiente no está dirigido específicamente a desestimar la felicidad como único fin último, sino que puede ser aplicado a cualquier intento de reducir la multiplicidad de fines últimos a uno sólo.

palda esta afirmación es, en palabras de Aristóteles, el siguiente:

"Tal parece ser eminentemente la felicidad [el más perfecto de los bienes], pues la elegimos siempre por ella misma y nunca por otra cosa, mientras que los honores, el placer, el entendimiento y toda virtud los deseamos ciertamente por sí mismos (pues aunque nada resultara de ellas, desearíamos todas estas cosas), pero también los deseamos en vista de la felicidad, pues creemos que seremos felices por medio de ellos."³

A primera vista, este argumento resulta sorprendente. ¿Cómo es posible que haya cosas que siempre deseamos por sí mismas y que sin embargo también las deseamos en virtud de otra cosa? Aristóteles sostiene los bienes pueden ser de tres tipos en este aspecto: por un lado están aquellas cosas que se buscan en virtud de alguna otra, y que constituyen los que hemos llamada fines intermedios o subordinados, por otro lado esta lo que siempre se elige por sí mismo y nunca por otra cosa, que son los que el supuesto exige del fin último y, por último, hay una tercera categoría de cosas que se eligen a la vez por sí mismos y por otra cosa, tales como los mencionados por Aristóteles en el párrafo citado arriba. Es importante subrayar que no son cosas a veces buscadas por sí mismas y a veces por otra cosa, sino que se buscan de ambas formas a la vez. Aristóteles supone que los bienes pertenecientes a esta última categoría tampoco pueden ser fin último, puesto que este ha de ser el bien más perfecto de todos que, sigue suponiendo, es el que se busca siempre por sí mismo y nunca por otra cosa. Los bienes pertenecientes al segundo grupo se buscan por sí mismos pero también y al mismo tiempo, en virtud de otra cosa, a saber, en virtud del fin último. Ahora bien, ¿cómo es esto posible? Es claro en que sentido podemos decir que tales cosas se buscan por sí mismas. El problema parece estar en que a la vez se busquen en virtud de otra cosa. ¿Qué puede querer decir esto? Podemos intentar contestar esta pregunta analizando la relación que tales cosas tienen con la felicidad.

La primera posibilidad es que cosas tales como los honores o la virtud sean causa de felicidad, del mismo modo que tomar un autobús es causa de que yo llegue a la Puerta del Sol. Pero entonces, parece que no son sino medios, o fines intermedios de otras cosas. Sin embargo, puede ser que, en un sentido, también y al mismo tiempo se busquen por sí mismas, a

³ Aristóteles, *Ética a Nicómaco*, 1097b.

saber, en el sentido de que las buscaríamos aunque nada se siguiera de ellas. Por ejemplo, podemos decir que tomar un autobús es un fin es sí mismo si, aunque sea también y al mismo tiempo deseado por conducirme a la Puerta del Sol. yo seguiría deseando tomar el autobús aunque no me llevara allí ni a ningún otro sitio. Del mismo modo, los honores serían un fin en sí mismos aunque su posesión no me diera la felicidad. Esta es una posible explicación de como es posible que algo sea buscado por sí mismo y a la vez en virtud de otra cosa. Pero en este caso las cosas que se buscan por sí mismas y a la vez en virtud de otra cosa tendrían el mismo estatuto que las cosas que se buscan sólo en virtud de sí mismas pues, en último término, también aquellas serían buscadas sólo por sí mismas en el caso de que no llevasen a nada, con lo cual tendríamos varios fines últimos y no uno sólo⁴.

Otra posibilidad es que los honores o el entendimiento se relacionen con la felicidad no como causa y efecto sino como las partes con el todo. Esta relación puede darse de tres formas. Supongamos que hay dos cosas, X e Y de las que puede decirse que se buscan por sí mismas y al mismo tiempo en virtud de la felicidad. Puede ocurrir, en *primer* lugar que la obtención de alguna de ellas o de ambas sea parte necesaria aunque no suficiente de la felicidad. En este caso, de nuevo, si las buscamos por sí mismas entonces las buscamos independientemente de que formen parte de la felicidad.

En *segundo* lugar, puede suceder que sean cada una o el conjunto de ambas partes suficientes aunque no necesarias de la felicidad, en cuyo caso la conclusión es la misma.

Por *último*, puede ocurrir que la obtención de estas cosas sea parte necesaria y que ambas juntas sean además suficientes. Pero entonces resulta que la obtención de esas cosas es la obtención de la felicidad. Ahora bien, pueden ocurrir dos cosas: o bien la felicidad es lo mismo que la obtención de ese conjunto de cosas o bien es algo distinto o, dicho de otro modo, puede ser que el todo sea o no igual a la suma de sus partes. Si lo es, entonces simplemente no tiene ningún sentido decir que estas

⁴ Es indiferente que pudiéramos suponer que cosas tales como los honores conducen irremisiblemente a la felicidad, pues aun entonces podríamos decir que en el hipotético caso de que esta no se siguiera de aquellas los seguiríamos buscando de cualquier forma, siendo esto requisito indispensable para que podamos decir que los honores se buscan por sí mismos.

cosas se buscan por sí mismas y a la vez en virtud de la felicidad, puesto que en este caso "felicidad" no sería sino otro nombre de "x e y". Y si la felicidad es algo distinto a la obtención de x e y volvemos a tener que concluir que, si estas cosas se buscan por sí mismas, entonces se buscan con independencia de que formen parte necesaria y suficiente de la felicidad.

En resumen, decir que x es buscado por sí mismo y es a la vez buscado por otra cosa, z, sólo tiene sentido si x y z son cosas distintas. Pero si son cosas distintas, entonces x sólo puede ser buscado por sí mismo si, a pesar de que también es buscado por z, sería buscado aunque no condujese a z.

La reducción substantiva puede aun llevarse a cabo de otro modo. Puede afirmar que sólo existe un fin último y al mismo tiempo explicar el hecho aparente de que hay diversas cosas que buscamos por sí mismas suponiendo que estos fines últimos aparentes sólo son tales en tanto que el único y auténtico fin último se encuentra en ellas como una parte en todo, es decir, realizando un movimiento inverso al anterior: no es que x e y formen parte del fin último, sino que el fin último es una parte de x e y. La dificultad ahora es que las cosas que buscamos por sí mismas no parecen tener nada en común. No sólo los honores y la virtud se buscan por sí mismos, sino también otras muchas cosas como ver una película o tomar un baño de sales o bailar. Lo único que cosas tan diversas como estas parecen tener en común es el hecho de que las deseamos. Podría entonces pensarse que tienen en común aquello por lo que son deseadas.

Quizá la respuesta más plausible es la ofrecida por el hedonismo, según el cual todo lo que es deseado es deseado en virtud del placer que contiene. Esto parece aplicarse directamente a cosas tales como el baño de sales, y otras cosas que producen algún tipo de placer corporal. Ahora bien, los hedonistas, por lo general, mantienen que este tipo de placeres no son los únicos intrínsecamente deseables. Además de los placeres corporales existen otros ligados al ejercicio de las facultades superiores del hombre. Así por ejemplo, ver una determinada película puede ser un placer de tipo intelectual, recibir honores, un placer de tipo social, ser virtuoso un placer moral y bailar un placer estético. Ahora bien, una vez que hemos admitido la existencia de tan diversos tipos de placer, empieza a ser dudoso que estos tengan en común algo más que el nombre.

Por otro lado, también parece dudoso que cosas que no tengan nada en común compartan el mismo nombre. De hecho, estas cosas comparten una característica que es la que permite que utilicemos en todos los casos el nombre de "placer" y que, por ello, es la que presta plausibilidad a la tesis hedonista, a saber, que de todas estas cosas se sigue lo que podríamos llamar "satisfacción".

¿Podemos decir entonces que buscamos todas estas cosas porque nos satisfacen? Pensemos en el baile. Cuando yo era pequeña, me madre me llevo a una academia de baile. Bailar me gustaba, me producía un cierto tipo de satisfacción. A otras niñas no les satisfacía, y dejaron de bailar tan pronto como pudieron. Si mañana dejara de satisfacerme bailar, dejaría de hacerlo. Parece entonces que podemos decir que bailo porque me satisface. Pensemos ahora en la virtud. En este caso, yo no busco la virtud porque me satisfaga ser virtuosa, sino por otros motivos. Cuando tengo un comportamiento virtuoso, esto me produce una cierta satisfacción, pero me satisface porque, en general, conseguir lo que se busca produce satisfacción. Y, precisamente por ello, siempre es posible encontrar el elemento común de la satisfacción asociado a la consecución de todo aquello que buscamos, aunque esta satisfacción no sea más que un bien sobrevenido y no motivo de la búsqueda. No es cierto entonces que busquemos las cosas porque nos satisfacen. Algunas cosas son buscadas por la satisfacción que nos produce, la consecución de otras nos produce satisfacción porque las buscamos.

No sólo la satisfacción por la cual buscamos determinadas cosas es distinta en cada caso, sino que en algunos casos la satisfacción no es causa del deseo de ciertas cosas, sino producto de su consecución. Todas esas cosas que deseamos no tienen otra cosa en común que el hecho de ser deseadas por nosotros. La tesis hedonista es plausible porque el cumplimiento de un deseo siempre produce una cierta satisfacción. Pero esto ni significa que esa satisfacción sea igual ni suficientemente similar en todos los casos, ni que sea la causa por la que deseamos todas las cosas³.

³ No sólo entran en esta categoría los casos relacionados con la moral. Por ejemplo, es bien posible que yo desee bailar por que deseo alcanzar la perfección física. Alcanzarla me satisface, pero no deseo alcanzarla por eso sino por otra cosa bien distinta.

Tal y como señalamos, todo lo dicho hasta ahora sirve tanto para la versión interpersonal como para la versión intrapersonal del supuesto según el cual algo es un fin último si y sólo si cumple con la condición de no ser *nunca* buscado por otra cosa y siempre por sí mismo. Hemos analizado las razones que existen para rechazar la teoría hedonista, que, debido a su plausibilidad, hemos utilizado como ejemplo del intento de encontrar un único fin último. Sin embargo, existe una razón de carácter general para rechazar este supuesto, al que llamaremos en adelante "supuesto débil". El supuesto débil se limita a exigir que el fin último no sea nunca buscado por otra cosa. Podemos entonces entender que es suficiente que algo no sea *de hecho* buscado más que por sí mismo. Sin embargo, que algo no sea de hecho buscado por otra cosa no significa que no pueda serlo. Es posible que alguien busque el placer porque le libere de tensiones, facilitando así la creación artística, que alguien busque la felicidad porque le hace ser más humanitario o que alguien busque la virtud porque sólo la virtud le puede garantizar la felicidad. En general, podemos suponer que todo lo que puede ser querido por otra cosa es de hecho, en algún momento querido por otra cosa⁴. Pero entonces, resultaría que algo que es ahora un fin último puede dejar de serlo dentro de un tiempo, al dejar que cumplir con el requisito expresado en el supuesto débil. Incluso aunque de hecho haya cosas que nunca lleguen a ser deseadas por otra cosa, la mera posibilidad teórica de que puedan llegar a serlo es suficiente para hacer ver lo absurdo de este supuesto.

Parece por tanto que, si queremos imponer sobre el fin último una condición plausible debemos defender lo que llamaremos "supuesto fuerte" según el cual algo es un fin último si y sólo si no puede ser deseado por otra cosa, es decir, según el cual lo característico y definitorio del fin último no es que nadie lo busque como medio sino que no pueda ser buscado de esta forma. Esta condición se impondría como única forma de garantizar que algo *nunca* sea querido por otra cosa, pues la única forma de garantizar que algo no suceda es que no pueda suceder.

Ahora bien, ¿qué debe ocurrir para que algo no pueda ser buscado

⁴ Este supuesto me parece tremendamente razonable, pues, dada la increíble variedad del ser humano, es de esperar que, con sólo suponer un espacio de tiempo suficiente, se satisfaga el supuesto. De hecho, hasta la fecha una increíble cantidad de las cosas más insospechadas han sido deseadas en virtud de las cosas más increíbles.

como medio? Hay dos alternativas: o bien se trata de algo que no puede tener ningún valor como medio, o bien se trata de algo máximamente valioso. Consideremos la *primera* alternativa. Piénsese que no se trata sólo de algo que no tenga de hecho ningún valor como medio, sino de algo que no pueda tenerlo, ya que de otro modo nos volveríamos a encontrar con las mismas dificultades planteadas por el supuesto débil. Dado que la cuestión de si algo tiene o no valor como medio es una cuestión empírica, es difícil entender que puede querer decir que algo no pueda tener valor como medio. De hecho, todos los ejemplos que se me ocurren son de cosas que tiene valor como medio en determinadas circunstancias y bajo ciertos supuestos o que pueden llegar a tenerlo. Sólo hay una forma en la que puedo pensar que algo es tal que no puede tener valor como medio, a saber, que no pueda ser un medio en absoluto, es decir, que sea tal que a) no se pueda seguir nada de ella en sentido causal y b) que no pueda formar parte real o abstracta de un todo de tal modo que pueda ser querida como medio para la obtención de este todo o de una parte inseparable de él. Una cosa tal no podría ser un medio en absoluto, de forma que, en caso de ser querida, sólo podría serlo en virtud de ella misma. Es difícil mostrar que tal cosa pueda existir, especialmente en lo que respecta a la condición a. En todo caso, la carga de la prueba recae en quien defienda la existencia de una cosa tal.

La *segunda* alternativa acerca de lo que debe ocurrir para que algo no pueda ser buscado como medio es que se trate de cosas tales que, si bien son o podrían ser medios y si bien tienen o podrían tener valor como tales, resulta absurdo pensar que puedan ser queridas como medios de otra cosa y no en virtud de ellas mismas dado que son máximamente valiosas. Pensemos por ejemplo en el conocimiento. El conocimiento puede ser medio para otras muchas cosas, por ejemplo para conseguir un buen puesto de trabajo. Puesto que puede ser un medio puede tener valor como medio, pero sería absurdo pensar que puede quererse únicamente en tanto que medio, pues esto significaría querer lo más valioso por lo menos valioso. Como máximo, puede suceder que se quiera *también* como medio, siempre y cuando su valor básico sea el que tiene por sí mismo. Así, aunque por medio del conocimiento puedan quererse matar dos pájaros de un tiro, habría que mantener que seguiría siendo deseado aun cuando no fuera un medio y no se siguiera nada de él.



Esta segunda alternativa es más prometedora que la primera, pero no parece llevarnos tampoco demasiado lejos. La razón fundamental es que entonces necesitaríamos a) una teoría acerca del valor que nos dijera qué es lo valioso y lo máximamente valioso, para poder localizar aquello que es un fin último y b) postular que lo que es máximamente valioso es de hecho siempre tenido por tal, ya que de lo contrario no tendríamos un fin último sino algo que debería serlo. Es decir, en ausencia de este segundo paso no habríamos demostrado que hay algo que no puede ser deseado más que por sí mismo, sino que hay algo que no debería ser deseado más que por sí mismo. De nuevo, la carga de la prueba recae sobre el que mantenga la posibilidad de formular una teoría satisfactoria en este sentido y de poder mostrar que lo máximamente valioso es de hecho fin último.

En resumen, hemos visto que el intento de reducir la aparente multiplicidad a nivel de fines intermedios a un único fin último se da en dos pasos:

1- en primer lugar, lo que llamamos reducción aristotélica muestra que debe existir algo que sea fin último pero que, sin embargo, no muestra que el fin último haya de ser único. Un supuesto no requerido por el argumento aristotélico pero relacionado con el y defendido explícitamente por Aristóteles da un paso más en este camino al exigir que el fin último no sólo sea aquello que sólo es querido en virtud de sí mismo, sino que sea aquello que *siempre* es querido en virtud de sí mismo y *nunca* por otra cosa.

2- el siguiente paso es dado por lo que llamamos reducción sustantiva. Esta reducción se plantea como una cuestión empírica, defendiendo que de hecho sólo hay una cosa que cumpla la exigencia de supuesto. Esta reducción podía seguir dos caminos distintos:

a) por un lado, puede argumentarse que las distintas cosas que aparentemente se buscan siempre por sí mismas se busca además y al mismo tiempo en virtud de otra cosa que sería el auténtico fin último

b) por otro lado, puede intentar mostrarse que lo que buscamos la aparente multiplicidad de fines es algo singular y común a todos ellos que constituye el fin último.

Hemos visto que ninguna de estas alternativas resultaba satisfactoria, y que, además, utilizan el mencionado supuesto en su versión débil que en sí misma resulta insatisfactoria. El supuesto sólo tiene sentido

si es utilizado en su versión fuerte, según la cual el fin último ha de ser tal que no pueda ser deseado más que en virtud de sí mismo.

También vimos como la exigencia contenida en el supuesto fuerte podía ser satisfecha de dos modos: o bien el fin último es tal que no puede tener valor como medio o bien es algo que resulta ser supremamente valioso, y vimos que ninguna de estas alternativas resultaba satisfactoria.

Por todo ello, podemos afirmar que la multiplicidad de fines que se da a nivel de fines intermedios también se encuentra a nivel de fines últimos. En primer lugar, porque ninguna de las reducciones examinadas resulta satisfactoria y, en segundo lugar, porque, en cualquier caso, ni el supuesto débil ni el fuerte bastan para concluir que el fin último ha de ser único, pues es posible que haya más de una cosa que cumpla con las condiciones impuestas en estos supuestos.

APENDICE 2

La aciclicidad queda definida por Sen del siguiente modo: la relación R es acíclica en el conjunto X si y sólo si

$$\bigwedge x_1, \dots, x_j \in X : [(x_1 P x_2 \ \& \ x_2 P x_3 \ \& \ \dots \ \& \ x_{j-1} P x_j) \longrightarrow x_1 R x_j]$$

donde P es la relación de preferencia estricta y R la relación "preferido o indiferente a"¹.

La aciclicidad evita las preferencias cíclicas en el sentido de que evita que la preferencia colocada jerárquicamente en el último puesto sea preferida a la que aparece en primer lugar, estableciendo que la preferencia que encabeza el orden jerárquico debe ser preferida o indiferente (a lo sumo) a la última. Sen demuestra que, al evitar las preferencias cíclicas, la condición de aciclicidad, junto con las otras dos condiciones (reflexividad y completud) que caracterizan una ordenación, es suficiente para garantizar la existencia de una función de elección. En efecto, la condición necesaria para que exista una función de elección definida para el conjunto X es que para todo subconjunto propio S de X haya uno o varios mejores elementos de S . Si el conjunto X violara la aciclicidad, entonces habría al menos un subconjunto S de X , formado por x_1, \dots, x_j elementos, tal que $x_1 P x_2 \ \& \ x_2 P x_3 \ \& \ \dots \ \& \ x_{j-1} P x_j$ y no sucede que $x_1 R x_j$.

Ahora bien, siendo R una relación de preferencia débil, $\neg(x_1 R x_j)$ implica que ni x_1 es preferido estrictamente a x_j , ni x_1 es indiferente a x_j . Y, puesto que R es completa, no queda más alternativa que la de que x_j sea preferido estrictamente a x_1 ².

Con esto queda formado un círculo entre los elementos de S , de modo que es imposible que haya un mejor elemento entre ellos, puesto que un mejor elemento es aquel que es preferido o al menos indiferente a todos los demás. En nuestro ejemplo, x_1 no es un mejor elemento puesto que x_j es preferido a él, pero a su vez x_j tampoco es un mejor elemento, ya

¹ (Sen 76, p.32)

² Puesto que la completud exige que $x_1 R x_j$ o que $x_j R x_1$, y por hipótesis no sucede $x_1 R x_j$, entonces se da necesariamente $x_j R x_1$. Ahora bien, $x_1 R x_j$ y $\neg(x_1 R x_j)$ es la definición de preferencia estricta, mientras que la definición de indiferencia requiere $x_1 R x_j$ y $x_j R x_1$, y no dándose uno de los miembros puede concluirse que $\neg(x_1 I x_j)$.

que x_{j-1} es preferido a él. Por tanto, la función de elección no estaría definida, puesto que fracasaría en encontrar un mejor elemento para este subconjunto S.

Si tenemos una función de elección, parece que hemos encontrado la solución a nuestro problema inicial. Recordemos cuál era este problema. Puesto que un agente tiene no sólo una preferencia sino un conjunto de preferencias, necesitábamos que este conjunto tuviera una determinada estructura que nos permitiera elegir entre ellas de un modo racional, e.d., elegir la alternativa que se prefiera en más alto grado. Tal estructura es una ordenación, a la cual exigimos que sea reflexiva, completa y transitiva. Sin embargo, parece que sustituyendo la transitividad por una propiedad más débil, la aciclicidad, tenemos el conjunto de preferencias estructurado de un modo que permite una elección racional entre sus miembros.

En efecto, esto es lo que hace la función de elección. Podemos ver esto aplicando la función de elección a nuestro ejemplo. Para ello vamos a ampliar nuestro conjunto de preferencias con la inclusión de una nueva alternativa. Supongamos que nuestro conjunto de preferencias X es: a-ir al cine, b-irme a la cama, c-ir al teatro y d-quedarme en casa viendo la televisión.

Supongamos que se da una relación R entre los miembros de este conjunto. Sea R: ((a,a), (b,b), (c,c), (d,d), (a,b), (a,c), (a,d), (b,c), (b,d), (c,a), (c,d), (d,a)).

La relación R es reflexiva, pues cada elemento está relacionado consigo mismo, es completa, pues cada uno de los elementos está relacionado con todos los otros. R también es acíclica, pues se cumple que si tomamos cualquier subconjunto de elementos, por ejemplo el subconjunto {a,c,d} y sucede que aPc y cPd , como es el caso, pues se dan los pares (a,c) y (c,d), entonces también sucede aRd . Y en nuestra relación se da aRd , pues se dan los pares (a,d) y (d,a)³. Naturalmente, para que se cumpla aRd no es necesario que se den ambos pares. Bastaría con que apareciera el par (a,d). Ahora bien, si apareciera sólo este par, entonces R

³ Se demuestra que la aciclicidad en los términos de un conjunto no implica aciclicidad en el conjunto (Sen 76,p.33). Esto no es importante en nuestro caso, pues R es aquí acíclica para todo el conjunto ya que se da aPb , bPc , cPd y aRd .

no sólo sería acíclica, sino también transitiva. Por eso hemos incluido ambos pares, puesto que lo que queremos ver es cómo puede darse elección racional con una condición más débil que la transitividad.

Es ahora fácil comprobar que en cada subconjunto de nuestro conjunto de preferencias existe un mejor elemento⁴. Por ejemplo, en el subconjunto $\{b, c, d\}$ el mejor elemento sería b . Lo mismo sucede en cada subconjunto propio y en el conjunto. El mejor elemento del conjunto sería a . Y puesto que hay un mejor elemento para cada subconjunto no vacío de nuestro conjunto de preferencias, hay una función de elección.

Tenemos por tanto un conjunto con una estructura más débil que la ordenación y que nos permite elegir un miembro del conjunto. En nuestro ejemplo, la función de elección seleccionaría al mejor elemento, a . Si este conjunto de preferencias tuviera la estructura de una ordenación, también existiría una función de elección y, naturalmente, el elemento seleccionado seguiría siendo a . Quizá se piense que en una estructura que sólo sea acíclica puede surgir el problema de que exista no sólo un único mejor elemento, sino varios, en cuyo caso la función de elección no señalaría inequívocamente a un sólo miembro. Sin embargo esto no es una objeción, a menos que lo sea también para una ordenación, pues en una ordenación definida sobre la relación R también puede suceder que se den varios mejores elementos, e.d., que a la cabeza de la ordenación aparezcan varias preferencias empatadas.

Habíamos dicho que la elección derivada de una ordenación era una elección racional. El motivo era que la ordenación establecía una jerarquía entre las alternativas de tal modo que se podía seleccionar la que tuviera un mayor grado de preferencia. Esto es posible puesto que una ordenación nos permite introducir una medida sobre las preferencias, la utilidad. A continuación, identificábamos la acción racional con la maximización de la utilidad.

Es evidente por qué pedíamos que la estructura del conjunto de preferencias fuera una ordenación y, por tanto, que fuera transitiva. Supongamos que el conjunto de las alternativas de nuestro ejemplo tiene

⁴ La comprobación de la aciclicidad sigue un procedimiento constructivo. Dado que X es finito, podemos examinar exhaustivamente el conjunto partes de X , comprobando que cada uno de sus miembros cumple esta propiedad.

la estructura de una ordenación. Esto quiere decir que R se definiría sobre ese conjunto mediante los mismos pares que definían la relación acíclica anterior, con la exclusión del par (d,a) . Por lo tanto, las alternativas aparecerían en el siguiente orden jerárquico: a,b,c,d , de tal modo que la alternativa que aparece en un puesto superior en la lista es estrictamente preferida a todas las demás (esto sucedería incluso si un mismo puesto, por ejemplo el segundo, estuviera ocupado por más de una alternativa). Esto es una condición previa para introducir una medida sobre las preferencias. Una vez introducida esta medida, diremos que cada alternativa tiene más utilidad que las que ocupan un puesto inferior a ella en el orden jerárquico. Y dijimos que una acción racional es la que hace máxima la utilidad, e.d., la que elige la acción a .

Es también inmediato ver que esto no es posible si la estructura no es transitiva sino sólo acíclica. En este caso si cada alternativa es estrictamente preferida a las siguientes, puede suceder que la última sea indiferente a la primera. Aquí no puede introducirse ninguna unidad de medida, pues, si intentáramos hacerlo, veríamos que a y d tienen la misma utilidad, que esta utilidad es superior a la de b , la de b a la de c y esta a la de d . Con lo cual resultaría que la utilidad de d sería al mismo tiempo la mayor y la menor de la lista, lo cual resulta absurdo para una medida.

Desde luego, si no puede introducirse una medida no puede identificarse la racionalidad práctica con la maximización de esa medida. Ahora bien, ¿es necesario introducir esa medida y, por tanto, realizar esa identificación?. Para contestar esta pregunta es necesario establecer una distinción entre medida ordinal y medida cardinal. Al introducir el concepto de utilidad diciendo que cada alternativa de la lista tiene más utilidad que las que aparecen por debajo de ella nos referíamos a la utilidad en tanto que medida ordinal. Decir que una alternativa tiene más utilidad que otra es decir que es preferida a ella. Pero no se dice nada respecto a la intensidad con la que es preferida.

Si nunca fuera preciso utilizar la medida cardinal podríamos argumentar que la medida ordinal tampoco es necesaria, pues el concepto de mejor alternativa supliría su función de señalar la alternativa que es al menos tan buena como todas las demás. En este sentido, sería lo mismo decir que una alternativa es la mejor que decir que es la que tiene mayor

utilidad.

El problema surge cuando en ciertas situaciones en las que puede (y suele) encontrarse el agente al elegir parecen requerir la utilización de una medida cardinal, con la que se definiría trivialmente la medida ordinal. Y en tales situaciones sería necesario identificar la racionalidad práctica con la maximización de la utilidad.

Más adelante estudiaremos con algún detenimiento estas situaciones, a saber, situaciones de riesgo e incertidumbre, y podremos ver si en efecto es necesario utilizar una medida cardinal.

Si la medida cardinal resulta necesaria, desde luego habrá que exigir que el conjunto de preferencias tenga la estructura de una ordenación y por tanto deberá que exigirse transitividad. Sin embargo, incluso ahora puede decirse algo a favor de la transitividad. Volvamos para ello a nuestro ejemplo. Si mis preferencias son acíclicas pero no transitivas, resulta que yo prefiero ir al cine a irme a dormir, prefiero ir a dormir que ir al teatro, ir al teatro a ver la televisión y me da lo mismo ir al cine que ver la televisión. No puede negarse que esta estructura de preferencias es bastante extraordinaria. De hecho, es tan extraordinaria que, con palabras de Gauthier "al desechar estas combinaciones de preferencias, la transitividad expresa nuestro ideal preteórico de la racionalidad de la elección" (Gauthier 86, p.41).

En efecto, lo que la aciclicidad pide es únicamente que la relación de preferencia estricta no se invierta en una serie, ya que esto daría lugar a la aparición de un círculo. Sin embargo, permite que la preferencia estricta entre los miembros de un conjunto ordenado genéricamente se resuelva en indiferencia entre el último elemento y el primero. Por otro lado, la idea preteórica no sólo pide que la relación de preferencia estricta no se invierta, sino que también pide que se mantenga, esto es, que sea una relación transitiva⁸.

⁸ La exigencia de transitividad no se mantiene para la relación de indiferencia. Como argumenta Gauthier (86,p.41) se puede ser indiferente entre cosas suficientemente similares, pero una sucesión de pequeñas diferencias puede producir una diferencia grande, de tal modo que ya no se sea indiferente entre dos miembros suficientemente alejados de la serie.

APENDICE 3

Tal y como hemos visto, Slote plantea un caso en el que alguien prefiere lo que él mismo piensa que es peor bajo todo punto de vista. ¿Como es esto posible? Lo que sucede es que tiene lo que Slote llama un "hábito de moderación". Esta moderación, sin embargo, no se corresponde con la virtud tradicional defendida, por ejemplo, por los epicúreos. Estos proponían la moderación en la persecución del placer como un buen medio para llevar una vida más placentera. La moderación de la que habla Slote no es una virtud instrumental. Pero tampoco es algo que haya de ser buscado por sí mismo "si con eso se quiere decir que es algún tipo de característica admirable o de virtud"¹. Es, simplemente, un hábito que algunas personas tienen. Y tenerlo "no es algo irracional ni estúpido, incluso si se reconoce que el hábito contrario de maximizar tampoco lo es". Para Slote, un individuo con tal hábito de moderación representa un modelo alternativo de racionalidad práctica igualmente válido que el representado por la maximización. Es el modelo de la satisfacción².

Es importante notar que esta preferencia por lo peor no se atribuye a alguna característica innata del individuo. Es más, se habla de esta tipo de preferencias como algo que surge de un determinado hábito. Esto es extremadamente curioso, si tenemos en cuenta que un hábito es algo que se adquiere mediante la práctica y, típicamente, con no pocos trabajos y gasto de energía por nuestra parte, sobre todo si tenemos en cuenta que no se trata de habituarse a lo bueno (cosa que, según el dicho popular, se hace pronto y fácilmente) sino a lo peor. Todos nosotros tenemos hábitos, algunos de los cuales hemos adquirido consciente y voluntariamente, mediante procesos que van desde lo simplemente molesto a lo auténticamen-

¹ Slote (85), p.40.

² No se debe confundir este modelo defendido por Slote con el propuesto por algunos economistas tales como H. Simon. El modelo económico de satisfacción surge para hacer frente a las condiciones de elección reales a las que se enfrentan los agentes individuales o colectivos y cuyas condiciones de aplicación, así como su justificación, ya hemos analizado (ver cap.3). Slote da un paso más que resulta ser definitivo para el asunto que ahora nos ocupa. El modelo económico no implica que sea racional rechazar algo mejor y preferir lo peor cuando en situaciones en las que ambas cosas están a la mano. Por ejemplo, según el modelo económico de satisfacción, el hombre de nuestro ejemplo se está comportando de un modo irracional. Sin embargo, el modelo de Slote si afirma que es racional rechazar lo mejor en tales circunstancias, afirmación que marca toda la diferencia.

te trabajoso. Por ejemplo, yo me he habituado a calentar los músculos antes de tomar una clase de baile. Hacerlo no es especialmente divertido, la mayoría de las veces no apetece en absoluto, es pesado, rutinario y nada creativo. Es, en suma, una pesadez. Y, por ello, me ha costado bastante adquirir el hábito. Ahora que ya lo tengo, me cuesta menos calentar diariamente. Pero el proceso fue costoso. Ahora bien, yo tenía una buena razón para someterme a mi misma a ese entrenamiento. Calentar antes de la clase evita lesiones y aumenta el rendimiento. Por eso, en algún sentido relevante no me he habituado a lo peor, sino a lo mejor. Sin embargo, el individuo del que habla Slote ha debido pasar por un proceso igual (¡más!) de trabajoso y difícil para llegar a adquirir el hábito de preferir lo peor. Y esto, además, es algo que se supone que muchos de nosotros hemos hecho.

No vamos a entrar ahora en si lo han hecho muchos, pocos o ninguno. Más interesante es saber si adquirir tal hábito es racional. Que lo sean (siempre según Slote) las acciones que de él se siguen no significa que lo sea adquirirlo. Pensemos en un drogadicto que esta atravesando una crisis de abstinencia y traslademos al caso de la racionalidad a la responsabilidad moral. Supongamos que este individuo comete alguna falta, como un pequeño robo mediante el que espera obtener dinero para una nueva dosis. No es moralmente responsable de sus actos en el estado en el que se encuentra. Pero si que lo es de haber adquirido el hábito del cual surgen tales acciones. Y hay otros muchos casos en los que lo que se predica del acto no se predica necesariamente del hábito que los produce³. Podemos decir, aunque esta formulación sea muy imprecisa, que un hábito es racional, o mejor aun, que es racional adquirir un hábito si tenemos un buen motivo para hacerlo, es decir, si el trabajo utilizado en adquirirlo se ve compensado por los beneficios que obtendremos de su adquisición. Esto es lo que sucede, por ejemplo, con el hábito de calentar los músculos antes de bailar y, en general, con todos los hábitos que adquirimos conscientemente. Sin embargo, este no es el caso del hábito de moderación del que habla Slote (en adelante "moderaciónS" por brevedad),

³ Un caso como el que nos ocupa pero inverso se analiza en el capítulo I, cuando nos ocupamos de la racionalidad indirecta, y vemos como en ciertas circunstancias puede haber actos irracionales que se sigan de la adquisición de un hábito o de una actitud que es racional adquirir.

aunque si lo es, por cierto, del hábito de moderación epicúrea. Podemos entonces decir que es irracional adquirir ese hábito en el sentido de que hacerlo supone invertir un considerable esfuerzo sin obtener beneficio alguno, e.d., en trabajar gratuitamente.

Tenemos pues unas acciones que surgen de un hábito que nadie tiene ninguna razón para adquirir, de un hábito irracional. Podemos ahora preguntarnos si tales acciones, e.d., las acciones satisfactorias son racionales. Tal y como vimos en el ejemplo del drogadicto, hay casos de acciones que resultan racionales una vez que uno tiene un hábito que es irracional adquirir. Imaginemos un hábito de moderación ligeramente distinto al de Slote. Mediante él, los beneficios que sobrepasan una determinada medida de satisfacción nos resultan indiferentes. Este hábito se diferencia del hábito de moderación habitual en que no se adquiere con vistas a los beneficios que a la larga nos proporciona su adquisición, motivo por el cual podemos decir que es un hábito irracional. Sin embargo, una vez que tenemos ese hábito, una acción satisfactoria es racional, o al menos tan racional como una acción maximizadora. Supongamos que el individuo que rechaza el dulce del ejemplo de Slote mencionado anteriormente tiene este hábito. Como está satisfecho, la degustación de un dulce no le produce ninguna satisfacción adicional, no obtiene de ella ningún beneficio. Suponiendo que la degustación del dulce es gratuita en todos los sentidos, entonces nuestro individuo es absolutamente indiferente entre tomar el dulce y no tomarlo. Ninguna de las dos cosas altera su nivel de beneficios. Por tanto, la acción satisfactoria en este caso es absolutamente racional. Pero no es que sea igual de racional que la acción maximizadora, sencillamente por que en este caso no hay acción maximizadora posible en absoluto. El individuo se encuentra instalado en su nivel de satisfacción máxima y ninguna acción posible es capaz de elevar este nivel. De hecho, este individuo, al rechazar el dulce, no está prefiriendo lo peor.

Un individuo que tiene el hábito de moderación^S se encuentra en un caso bien distinto. Tal hábito no hace que, tras alcanzar un cierto nivel de satisfacción, el individuo sea insensible a un incremento en el nivel de utilidad (e.d., que no haya para él incremento posible de utilidad). El individuo de Slote estaría mejor si tomara el dulce. El que este individuo esté satisfecho sólo quiere decir que "no está mal", pero, desde luego, podría estar mejor. Pero su hábito le hace rechazar lo mejor en

virtud de lo peor. Y en esta medida, no sólo la adquisición del hábito es irracional, sino que también lo son las acciones que de él se siguen.

Por tanto, según nuestra argumentación, la acción satisfactoria defendida por Slote es irracional. Slote admite que la acción maximizadora es racional y se limita a añadir que la acción satisfactoria también lo es. Sin embargo, no ofrece nada que pueda ser considerado como un argumento en defensa de su postura. Parece que simplemente se limita a apelar a nuestra simpatía por la idea de la moderación⁴. Dejando aparte el hecho de que esta apelación no resulta una defensa convincente, es también dudoso que tengamos tal simpatía. La aparente efectividad de la apelación de Slote se basa en dos confusiones relacionadas entre sí. La primera confusión surge del propio término "moderación" que no se asocia habitualmente con el concepto de Slote, sino con la idea de raíces epicúreas. Puesto que son dos conceptos claramente distintos, no conviene traspasar la posible simpatía hacia uno de ellos al otro. De hecho, la simpatía por el concepto clásico es, en todo caso, un apoyo a la idea de maximización, cuya persecución es el mejor argumento a favor de la práctica de la moderación.

La segunda confusión, relacionada con la anterior, alude a los objetos típicos de la moderación. En efecto, la idea de moderación está asociada fundamentalmente a los bienes a) cuya persecución immoderada puede resultar contraproducente y b) de los que es posible tener "demasiado". El ejemplo de Slote mencionado alude de hecho a uno de esos bienes, la comida. Cuando se trata de ellos, siempre es posible encontrar razones a favor de la moderación. Es precisamente el apelar a este tipo de casos lo que da cierta verosimilitud a las afirmaciones de Slote. Pero pensemos en otro tipo de bienes, como por ejemplo la salud. Si bien puede sonar sensato el rechazo de un dulce porque uno ha comido "suficiente", es difícil defender la racionalidad de alguien que rechaza un incremento de su salud porque considera que ya está "suficientemente" sano. Naturalmente, se puede (y a veces se debe) ser moderado en la persecución de la salud, pero no porque sea racional conformarse con un cierto nivel, sino porque su búsqueda desmedida puede resultar insana a la larga o porque puede implicar la disminución de otros bienes. Es decir, en el

⁴ Slote (1985), p. 44

ejemplo de la salud se ve claramente que, en cualquier caso, la moderación es defendible por razones que tienen que ver con la maximización y no con la satisfacción, como es el caso de la moderaciónS.

Hemos visto las razones para rechazar el modelo de satisfacción propuesto por Slote. Podemos ahora concluir que si, tal y como hemos defendido, la conducta racional es la conducta maximizadora, entonces la conducta satisfactoria resulta irracional por consistir en la elección de lo peor.

APENDICE 4.

Supongamos que a es un dólar, b un centavo y c la muerte. Podemos asumir sin arriesgarnos que estos tres resultados cumplen con la condición de que a es preferido a b y b a c . Ahora bien, si aplicamos la condición de continuidad nos encontramos con que hay una probabilidad p tal que somos indiferentes ante la alternativa de obtener un centavo con seguridad y una apuesta que nos ofrece la probabilidad p de ganar un dólar y la probabilidad $1-p$ de morirnos. Pero, argumenta Gauthier, esto es absurdo, porque ¿estaría yo dispuesto a correr el riesgo de morirme a cambio de un dólar si no me muero?. Si la respuesta es negativa, entonces resulta que yo prefiero un centavo a tal lotería, con lo cual se incumple la condición de continuidad.

Parece que el problema en este caso es la atribución de valores de utilidad a los tres resultados. Concretamente, el problema reside en que el valor de c parece ser muchísimo menor que el de a . Sin embargo, esto es solamente parte del problema. Consideremos un caso donde los valores de a y c difieran mucho. Por ejemplo, supongamos que el valor de a es de 100.00 y el de c cero. En este caso, si b vale 100 el valor de p tendría que ser 0.001 para que fuéramos indiferentes entre conseguir b con seguridad y una lotería entre a y c con las probabilidades respectivas de p y $1-p$. Podríamos aumentar aún más la diferencia y el resultado seguiría sin parecer absurdo.

Lo que hace que el ejemplo de Gauthier sea tan impactante es que el resultado c es la muerte. Esto se debe a que por lo general consideramos a este resultado tan malo que no puede haber otro peor por malo que sea¹.

Si la ordenación es completa cada resultado está relacionado por la relación de preferencia con el resto de los resultados alternativos. Esto no supone un problema, pues colocaríamos a c en el último puesto de la jerarquía. Esto es, su utilidad es menor que la de cualquier otra cosa. El problema surge cuando queremos asignarle un valor de utilidad pre-

¹ Digo por lo general porque, naturalmente, esto no es así necesariamente. Para algunas personas hay cosas peores que la muerte, como traicionar a su fe para los mártires o vivir sin su amor para Romeo y Julieta.

ciso. Parece difícil saber cuánto menos vale la muerte o la traición que un dólar.

Sin embargo, este aspecto intuitivo del problema puede resultar confundente. Quizá sea mucho más clarificador utilizar el cálculo. Si podemos asignar un valor a estos sucesos, entonces es posible encontrar un valor de p . La cuestión está en encontrar una asignación de valor para este tipo de resultados que refleje el hecho de que no sólo son mucho peores que ganar un dólar, sino infinitamente peores. Necesitamos un valor de c que sea infinitamente más bajo que los de a y b . Esto quedaría reflejado si suponemos que el valor de c tiende a menos infinito.

Lo importante es ver qué valor toma p bajo esta asignación de valor a c . En los ejemplos que hemos utilizado anteriormente (p.42) puede verse que el valor de p disminuye al aumentar el de a . Es decir, cuanto más grande es uno de los premios de la lotería más dispuestos estamos a pagar el precio del boleto (la seguridad de b) con una probabilidad menor de ganar, e.d., de conseguir a en vez de b . Esto sucede incluso si el valor de c es caro, lo que podemos interpretar como no ganar nada. Que esto suceda así es perfectamente intuitivo. De hecho, es lo que sucede en todas las apuestas que realizamos. En estos casos el valor de p puede ser interpretado como la mínima probabilidad de ganar racionalmente aceptable dado el posible premio, a , y el precio de la apuesta, b .

De igual modo que el valor de p disminuye al aumentar el de a , el valor de p crece al disminuir el valor de c . De nuevo, esto es perfectamente intuitivo, pues significa que cuanto peor es el peor resultado posible de la lotería mayor debe ser la posibilidad de ganar para que merezca la pena correr el riesgo.

¿Qué pasaría si el valor de c tendiera a menos infinito?. Como es de esperar, en estos casos el valor de p tiende a 1. Esto es tanto como decir que en los casos límites no habría apuesta.

Gauthier tiene razón al decir que en estos casos preferiríamos b , a menos que nos ofrecieran a cambio la seguridad de a . En lo que ya no está tan claro que tenga razón es al decir que esto hace que la condición de continuidad sea sospechosa y que, para hacerla aceptable, sea preciso limitar su aplicación excluyendo los casos extremos². No se entiende muy

² El propio Harsanyi, uno de los mayores defensores de la condición de continuidad, dice que la

bien qué quiere decir que la continuidad falla en estos casos. Lo que tiene que suceder es que sea siempre posible determinar un valor de p dados tres valores de utilidad y esto sucede para *cualquier* valor que las alternativas puedan tomar.

Ahora bien, existen casos límite. En estos casos cuanto más se aproxima a a infinito más se aproxima p a 0 y cuanto más se aproxima c a menos infinito más se aproxima p a 1. Pero también se puede determinar un valor para p . Lo que sucede es que si p es muy próximo a 0 o a 1 el juego se desvirtúa progresivamente al perder su carácter arriesgado, pero nunca se llega a un punto en el que no haya lotería posible porque p haya tomado los valores 0 o 1. El valor de p esta siempre en el intervalo 0-1. En los casos límite p tiende a 1 o a 0, pero nunca llega a tomar esos valores. Precisamente esto es lo que hace creíble el ejemplo de Gauthier, a saber, que aún en el caso de que la utilidad de c tienda a menos infinito, p tiene un valor muy alto, pero la lotería sigue siendo una lotería y, por lo tanto, arriesgada.

Gauthier dice que en estos casos violaríamos la condición de continuidad al preferir b . Esto es muy discutible. Es verdad que preferiríamos b a menos que p tuviera el valor 1. Pero si el valor de utilidad de c tiende a menos infinito, e.d., tiene un valor tan bajo como queramos y más bajo que ningún otro posible, entonces el valor de p sería tan cercano a 1 como quisiéramos, tanto que lo más probable es que en la práctica sería razonable tratarlo como si fuera 1.

No parece haber por tanto ningún motivo para pensar que la condición de continuidad falla, ni siquiera en los casos límite. Gauthier acepta esta condición a pasar de que le parezca dudosa su aplicación en casos límites debido a que con esas escasas excepciones la continuidad parece una condición sumamente razonable. Por lo tanto nuestra razón para aceptarla será aun mayor si consideramos que ni siquiera en los casos límite puede decirse que no es aplicable.

APENDICE 5

El método de asignación de utilidades cardinales es a grandes rasgos como sigue. Se asigna un valor arbitrario tanto a la alternativa más preferida como a la menos, habitualmente 1 y 0 respectivamente. Una vez fijada esta utilidad es posible determinar una medida cardinal de utilidad para todos los demás resultados alternativos. Llamemos X al resultado más preferido y Z al menos. La utilidad de cualquier otro resultado alternativo Y será igual a p , donde p es el valor de probabilidad que hace que el individuo sea indiferente entre la certeza de Y y una lotería con la probabilidad p de conseguir X y la probabilidad $1-p$ de conseguir Z.

Veamos un ejemplo. Supongamos que un individuo está dispuesto a pagar cinco pesetas por un boleto de lotería con el cual puede ganar mil pesetas con una probabilidad de 0.001 o no ganar nada con una probabilidad de 0.999. Sea a "ganar mil pesetas", b "no ganar ni perder nada" y c "perder cinco pesetas". Puesto que nuestro individuo está dispuesto a comprar el boleto, podemos decir que es indiferente entre la seguridad de b y la lotería $(pa + (1-p)c)$, e.d., para él la utilidad de b es igual a la utilidad esperada de esta lotería. La utilidad esperada de esta lotería $UE(L)$ es igual a la utilidad esperada de a, $UE(a)$, más la utilidad esperada de c, $UE(c)$. Sabemos que la utilidad esperada de un resultado es el producto de su utilidad inicial y las probabilidades de que se realice. De modo que $UE(a) = 0.001 \times 1$ y $UE(c) = 0.999 \times 0$, con lo cual $UE(L) = 0.001$. Por lo tanto, la utilidad de b será 0.001, e.d., el valor de p .

Este método de cardinalización fue originariamente formulado por von Neumann y Morgenstern, quienes demostraron que si un agente cumple con los postulados de racionalidad que hemos expuesto, entonces podemos utilizar este método de asignación de utilidad cardinal y el comportamiento de tal agente podrá interpretarse como un intento de maximizar la utilidad esperada.⁴

⁴ Estas funciones de utilidad utilizan 1 y 0 como los valores de referencia para definir a partir de ellos el valor de utilidad de cualquier otra alternativa. La elección de estos valores es arbitraria. Esta arbitrariedad no plantea problemas, al menos cuando se trata de asignar utilidades a las preferencias de un sólo individuo, pues cualquier transformación lineal de la función de utilidad que utiliza estos valores es también una función de utilidad equivalente, ya que de lo que se trata es de

La crítica más generalizada contra este método es que las funciones de utilidad de von Neumann-Morgenstern (funciones de utilidad vNM) miden no sólo la utilidad esperada de los resultados sino también la actitud hacia el riesgo del agente. Esto sucede porque, como hemos visto, estas funciones determinan el valor de utilidad de una alternativa b para un agente especificando la lotería que habría que ofrecerle a ese agente para que fuera indiferente entre esa lotería y la seguridad de b . Ahora bien, si un agente tiene una actitud negativa hacia el riesgo entonces es de esperar que la utilidad esperada de la lotería que se le ofrece ha de ser muy alta para que esté dispuesto a perder la seguridad de b , mientras que si otro agente tiene una actitud más positiva hacia el riesgo puede aceptar una lotería peor.

Los defensores del uso de estas funciones de utilidad suelen argumentar que este tipo de críticas se fundamenta en una mala interpretación de estas funciones. Por ejemplo, Harsanyi dice que si bien es cierto que estas funciones expresan la actitud hacia el riesgo del agente, no es cierto que lo que hacen sea registrar esas actitudes sino explicarlas en términos de la importancia relativa que tienen para el agente las posibles ganancias y pérdidas de la lotería (Harsanyi 77b,p.642). Es decir, si yo sólo estoy dispuesta a aceptar el boleto de lotería de nuestro ejemplo si se me ofrece una probabilidad de ganar de 0.005, la explicación que se da no es la de tomar la actitud hacia el riesgo como un dato primitivo y decir que mi actitud es más negativa que la del agente que acepta una probabilidad de 0.001, sino la de explicar esta actitud en términos de la utilidad de las distintas alternativas. De este modo, lo que la función de utilidad vNM dice en este caso es que para mí cinco pesetas tienen una utilidad relativamente mayor que la que tienen para el otro agente. Concretamente, para mí cinco dólares tienen una utilidad de 0.005 y para él de 0.001. Esta diferencia en el valor de utilidad asignado a los cinco pesetas puede deberse a muchos factores. Por ejemplo, nuestro agente puede tener una deuda de mil pesetas que desea pagar, de

medir la utilidad relativa de una alternativa frente a otras, para lo cual lo importante es el intervalo (ver Harsanyi 77, p.40)

El motivo por el que resulta conveniente utilizar 1 y 0 como valores de referencia es que esto hace la función matemática más sencilla. En efecto, el valor de p es $b-c/a-c$, de modo que si $c=0$ entonces $p=b/a$, y si $a=1$ entonces $p=b$.

modo que está dispuesto a jugar una lotería muy mala que le ofrezca la posibilidad de ganar mil pesetas porque para él tener cinco pesetas no supone una solución a su problema. Por otro lado, yo puedo estar ahorrando para comprarle un regalo de diez pesetas a un amigo, de modo que sólo me jugaría mis ahorros de cinco pesetas si me ofrecieran una lotería mejor.

Esta interpretación de la actitud hacia el riesgo es bastante intuitiva. De hecho, el mejor modo de saber lo que algo vale para alguien es saber lo que estaría dispuesto a arriesgar para conseguirlo.

Por tanto, las funciones de utilidad VNM son perfectamente aplicables para los agentes cuya actitud hacia el riesgo dependa únicamente de las utilidades de las posibles ganancias o pérdidas.

El problema parece surgir cuando el agente tiene una actitud positiva o negativa hacia el riesgo mismo. Parece ser que en estos casos no está tan clara la aplicabilidad de estas funciones.

Las condiciones de racionalidad que hemos impuesto para que sea posible la aplicación de estas funciones desestiman tales actitudes hacia el riesgo. En especial este es el caso de la condición de monotonicidad. En ella se asegura la dependencia directa de la actitud hacia el riesgo respecto a los posibles premios, estableciendo que siempre se preferirá una lotería menos arriesgada si el valor de los premios es el mismo.

Sen plantea un ejemplo en el que la monotonicidad no se mantiene (Sen 76, p.122). En este ejemplo el agente es un montañero que ama el riesgo. Debido a este amor, el agente, si bien preferiría una probabilidad del 95% de sobrevivir a una del 80%, también preferiría una probabilidad del 95% a una de 100%. Bajo el supuesto lógico de que nuestro agente prefiere sobrevivir a no sobrevivir, resulta que nos encontramos con una violación de la monotonicidad.

Puesto que se viola una de las condiciones de racionalidad, no resulta problemático para nuestra teoría que en este caso no resulten aplicables las funciones VNM, pues estas se aplican sólo en los casos en los que el agente cumple con esas condiciones. Lo perturbador del ejemplo reside en que parece que es arbitrario desestimar a los amantes del riesgo como agentes racionales. En efecto, no parece haber razón alguna para afirmar que la conducta de un agente no es racional simplemente porque para él el riesgo no sea un mal necesario sino algo por lo que se expresa

una clara preferencia.

Sin embargo, esto es un error. El amor al riesgo puede tratarse como una preferencia más que entra en el cálculo junto con las otras preferencias sin que esto contradiga a la teoría. Veamos el ejemplo del montañero con más detenimiento.

Supongamos que este agente se enfrenta con las siguientes alternativas: a-vivir de una manera arriesgada (por ejemplo dedicándose al montañismo), b-vivir de una manera que no implique riesgos innecesarios (trabajando en un banco, por ejemplo, y dedicando el ocio a coleccionar sellos), c-morirse. Supongamos también que el agente prefiere a a b y b a c.

La teoría que estamos defendiendo no tiene ningún problema para tratar con estas preferencias. De hecho, nuestra teoría no dice nada acerca del contenido sustantivo de las preferencias. Lo único que exige es que se cumplan ciertas condiciones. Y estas preferencias pueden cumplirlas tan bien como cualesquiera otras.

Es presumible que nuestro montañero cumpla con la condición de monotonicidad. Es decir, es presumible que si se le ofreciera una probabilidad mayor de obtener el resultado deseado, a saber, vivir arriesgadamente, preferiría esta lotería a otra que le ofreciera el mismo resultado con una probabilidad menor. Si esto sucede así, entonces podríamos calcular el valor de b mediante la utilización de funciones vNM.

Nuestra teoría no exige que no se tengan actitudes hacia el riesgo por el riesgo mismo. Lo único que exige es que al menos en un caso la actitud hacia el riesgo esté determinada únicamente por la utilidad de los posibles premios de la lotería. Nuestro montañero violaría la monotonicidad si prefiriera una lotería del 95% de probabilidades a favor del resultado de vivir arriesgadamente a una del 100%. Pero si violara esta condición no tendríamos ningún inconveniente en aceptar que se esta comportando de un modo irracional.

Tenemos pues que las funciones vNM interpretan la actitud hacia el riesgo como un indicador de la relativa importancia que tienen para un agente los distintos premios. Hemos visto un ejemplo de este tratamiento de la actitud hacia el riesgo y argumentado que es una interpretación muy intuitiva. Sin embargo, los críticos de nuestra teoría aun podrían decir algo. Su argumento podría ser más o menos el siguiente. Supongamos que el

agente M está dispuesto a jugar cinco pesetas a la lotería y que el agente N no lo está. La interpretación de la teoría es que M asigna un valor relativo menor a las cinco pesetas comparados con las 1000 pesetas del premio que el valor relativo que N asigna a estas mismas cinco pesetas. Sin embargo, esta interpretación puede no ser acertada. Puede ser que lo que ocurre es que N piensa que el juego es pecaminoso o que propicia determinadas actitudes poco deseables en el jugador. En este caso, N no aceptaría ninguna lotería, con independencia de lo que supongan para él tanto el premio de la lotería como el precio de la misma. O quizá habría que tentarle mucho más para que jugara.

Esta crítica se diferencia de la anterior en que aquella se basaba en una mala interpretación de lo que significa la condición de monotonidad, mientras que esta tiene su punto de apoyo fundamental en una interpretación demasiado literal de la metáfora de las loterías. Esta metáfora surge del hecho de que, en situaciones de riesgo e incertidumbre, elegir una acción es como hacer una apuesta en el sentido de que no se sabe con certeza lo que va a resultar de la acción. También, igual que en las loterías reales podemos saber cómo valora un jugador el dinero de la apuesta sabiendo cómo tiene que ser la apuesta para que esté dispuesto a jugar, en las situaciones de riesgo e incertidumbre podemos saber cómo valora un agente uno de los posibles resultados si sabemos que combinación de resultados y probabilidades le harían indiferente a la seguridad de conseguir ese resultado.

Pero, si bien las apuestas son un tipo de comportamiento arriesgado o incierto, no todo comportamiento de este tipo es una apuesta en sentido literal. Pensemos en nuestro ejemplo del sábado por la noche. Tenemos cuatro alternativas que son posibles resultados de nuestras posibles acciones. Como se recordará, estos resultados son: a-ver "El último emperador", b-ver "Enrique IV", c-quedarme en casa y d-darme un paseo en balde. Supongamos que yo prefiero estos resultados en ese orden. Se puede saber qué valor de utilidad le asigno a la alternativa b averiguando qué probabilidad de ver la película y qué probabilidad de darme un paseo en balde me haría arriesgar la seguridad de b. De nuevo, esto es perfectamente intuitivo. En este caso, la cardinalización depende del uso de la hipótesis de una decisión arriesgada y en este sentido puede hablarse de lotería. Pero esta lotería no tiene nada que ver con el juego de los ca-

sinos ni con la lotería nacional, de modo que si lo que queremos determinar es la utilidad que tiene b para un miembro de la liga antijuego este no puede objetar que nuestra apuesta va en contra de sus ideales y que él prefiere la seguridad de b a cualquier combinación de resultados y probabilidades, porque el juego que utilizamos para determinar la utilidad no tiene nada que ver con el juego contra el que lucha la liga de la que es miembro. De hecho, cuando este agente se encuentra en una situación que no es de certidumbre (lo que presumiblemente le sucederá muy con mucha frecuencia) valora las probabilidades sin que esto le suponga un problema de conciencia.

APENDICE 6.

Podemos ver como funciona el principio de concesión de Zeuthen-Harsanyi con un ejemplo simple. Supongamos que dos personas tienen la oportunidad de cooperar en un negocio que producirá 100 pesetas (que tal ganancia no merezca ni siquiera la mínima molestia de iniciar un regateo no es relevante. Podría escogerse cualquier otra cifra superior y más realista, pero esto alargaría el ejemplo sin ofrecer ningún tipo de ventaja). El caso más sencillo es aquel en el que en caso de no cooperar ninguno de los dos participantes obtiene beneficios.

Al inicio de la negociación, cada una de las partes propone un acuerdo. Estas propuestas deben cumplir 3 condiciones:

-deben consistir en la realización de un resultado óptimo, pues de otro modo habría algún otro resultado posible que beneficiaría más a ambos, o al menos para uno sin perjuicio del otro.

-cada uno de los jugadores propondrá el resultado que le rinda a él mismo el máximo beneficio posible.

-cada propuesta realizada por uno de los jugadores debe ofrecer al otro al menos tanto como recibiría en caso de que no se llegara a un acuerdo. Esta condición es evidentemente necesaria, pues de otro modo el contrario no sólo no se beneficiaría de la cooperación, sino que perdería con ella, por lo que de ningún modo accedería a cooperar¹.

Llamaremos A_1 a la propuesta del jugador 1 y A_2 a la propuesta de 2. Cada una de estas propuestas consistirá en una utilidad U para el jugador 1 y una utilidad V para el jugador 2. Llamaremos (U_1, V_1) a la distribución de utilidades en la que consiste A_1 y (U_2, V_2) a la correspondiente de A_2 .

Supongamos que $(U_1, V_1) = (99, 1)$ y que $(U_2, V_2) = (1, 99)$. A partir de estas propuestas cada uno de los jugadores puede calcular lo que concedería en caso de aceptar la propuesta. Llamemos C_1 a la concesión del jugador 1 y C_2 a la de dos. Puede calcularse C_1 y C_2 del modo siguiente:

$$C_1 = 99 - 1/99 = 0,98 \qquad C_2 = 99 - 1/99 = 0,98$$

Puesto que el valor de la concesión de ambos es el mismo y es un valor

¹ Esta última condición no resulta importante en el caso concreto de nuestro ejemplo, ya que en caso de no cooperación ninguno conseguiría nada, pero es extremadamente importante en otros casos.

mayor que 0, ambos deben hacer una concesión simultánea. Esta concesión consistirá en una nueva propuesta que asigne una utilidad menor al jugador que la hace de la que le asignaba la propuesta anterior pero que sea aun mayor que la que le asigna la propuesta del contrario. Supongamos que 1 hace una propuesta A_3 tal que $(U_3, V_3) = (95, 5)$ y 2 hace una propuesta A_4 tal que $(U_4, V_4) = (30, 70)$. Entonces

$$C_1 = 95 - 30/95 = 0.68$$

$$C_2 = 70 - 5/70 = 0.92$$

puesto que $C_1 < C_2$, 1 debe hacer una concesión. Supongamos que 1 propone $(65, 35)$. Entonces

$$C_1 = 65 - 30/65 = 0.53$$

$$C_2 = 70 - 35/70 = 0.5$$

Ahora $C_2 < C_1$. 2 debe hacer una nueva propuesta que suponga una concesión, por ejemplo $(45, 55)$. Esto hace que

$$C_1 = 65 - 45/65 = 0.30$$

$$C_2 = 55 - 35/55 = 0.36$$

es decir, $C_1 < C_2$. 1 puede conceder ahora $(54, 46)$, con lo cual

$$C_1 = 54 - 45/54 = 0.1666$$

$$C_2 = 55 - 46/55 = 0.1636$$

esto hace conceder de nuevo a 2, ya que $C_2 < C_1$. 2 puede proponer $(49, 51)$, lo que haría

$$C_1 = 54 - 49/54 = 0.092$$

$$C_2 = 51 - 46/51 = 0.098$$

con lo cual obliga a 1 a hacer la siguiente concesión. La concesión mínima que 1 puede hacer para conseguir que el valor de su concesión sea superior al valor de la concesión de 2 es $(50, 50)$. Con esta propuesta

$$C_1 = 50 - 49/50 = 0.02$$

$$C_2 = 51 - 50/51 = 0.019$$

lo que hace $C_2 < C_1$. Si 2 intentara una propuesta que le diera a él mismo una utilidad algo mayor y a 1 algo menor que 50, por muy pequeña que fuera esta diferencia, el resultado sería que el valor de su concesión seguiría siendo menor al valor de la concesión de 1. Por ejemplo, si 2 propusiera $(49.99999, 50.00001)$, entonces

$$C_1 = 50 - 49.99999/50 = 0.000002$$

$$C_2 = 50.00001 - 50/50.00001 = 0.000001$$

Es decir, la única propuesta de 2 que podría hacer el valor de su concesión al menos igual que el valor de la concesión de 1 es $(50, 50)$, pues entonces

$$C_1 = 50 - 50/50 = 0$$

$$C_2 = 50 - 50/50 = 0$$

lo cual significa aceptar la propuesta de 1. Esta es la única solución viable para 2, ya que si intentara hacer que la concesión de 1 fuera menor que la suya debería proponer un resultado que le diera a él menos de 50 y a 1 más, lo cual significaría proponer un resultado más desfavorable

para él mismo que el resultado que propone el contrario. Por consiguiente, el acuerdo consistirá en la elección de una estrategia conjunta cuyo resultado sea (50,50).

Este ejemplo pone de relieve algunas características generales del principio de concesión de Zeuthen-Harsanyi. Una de estas características es que a lo largo del proceso de regateo el valor de las concesiones de ambos jugadores va acercándose progresivamente a cero. Cuando $C_1 = C_2 = 0$ la negociación ha llegado a su término. Esto no sucede en los casos en los que $C_1 = C_2 > 0$. En estos casos, la teoría de Zeuthen-Harsanyi exige que ambos jugadores hagan una concesión simultánea. La razón es que en estos últimos casos, los dos harían una concesión muy grande al aceptar la propuesta del otro. Como además ninguno tiene más razón que el otro para conceder, la negación se paralizaría sin haber llegado a un acuerdo cooperativo. Por el contrario, en los casos en que $C_1 = C_2 = 0$, se ha llegado a un punto en el que uno de los jugadores considera que lo más ventajoso para él es aceptar la propuesta del otro.

Esto nos lleva a la segunda característica. El procedimiento de Zeuthen-Harsanyi siempre culmina en un punto en el que uno de los jugadores acepta la propuesta del otro. En efecto, es fácil comprobar que este es el único medio de hacer que ambas concesiones tengan el mismo valor y que este sea igual a cero. Naturalmente esto puede suceder en cualquier momento de la negociación. Por ejemplo, cuando el jugador 1 ofrece (54,45), si 2 acepta entonces

$$C_1 = 54 - 54/54 = 0$$

$$C_2 = 45 - 45/45 = 0$$

Pero si esto sucediera, 2 estaría actuando irracionalmente, pues estaría aceptando un resultado inferior para él a otro resultado que podría conseguir. Esto sucede por que 2 puede forzar a 1 a hacer una nueva concesión, e.d., a ofrecer un acuerdo más ventajoso para 2, por ejemplo proponiendo el resultado (49,51) tal como sucede en nuestro ejemplo. Un jugador racional, por definición, sólo hará la concesión mínima suficiente para conseguir que el contrario ofrezca algo más ventajoso para él.

Cuando la negociación llega a un punto en el que uno de los jugadores tiene que aceptar la propuesta de otro, dando así por concluido al proceso de regateo al hacer que ambas concesiones sean iguales a cero, esto sucede por que al jugador que acepta la propuesta del otro *no le queda más remedio* que hacerlo así, e.d., esto es lo más ventajoso para el

pues, como vimos en nuestro ejemplo, cualquier otra propuesta por su parte que le fuera mínimamente más ventajosa que la que le ofrece el contrario seguiría haciendo el valor de su concesión menor que el de su oponente y para hacer que ese valor fuera más grande para el otro que para el tendría que proponer un trato que le fuera menos ventajoso que el propuesto por la parte contraria.

El último punto que conviene señalar es que el momento en el que a un jugador no le queda más alternativa que aceptar la propuesta de su contrario es precisamente el momento en que la última propuesta (U_n, V_n) es tal que $U_n \cdot V_n > U \cdot V$, siendo U y V cualquier utilidad de 1 y 2 respectivamente derivada cualquiera de los posibles términos del acuerdo de la situación de regateo de la que se trate. Y precisamente esta característica del principio de concesión de Zeuthen-Harsanyi es la que hace que este método para determinar qué punto óptimo será escogido como término del acuerdo cooperativo sea equivalente al de Nash. Esto es, es uno y el mismo punto el que en todos los casos i) maximiza el producto de las utilidades de las partes y ii) hace que el valor de la concesión de una parte sea igual al de la otra y ambos igual a cero.

En los casos en los que el *status quo* es distinto de cero para uno de los jugadores o para ambos, la aplicación del principio es ligeramente distinta, tanto en el caso de que U^* y V^* sean iguales como en el caso de que sean distintos. Esto es necesario desde el momento en que la solución al problema de regateo debe tomar en cuenta las utilidades que los jugadores conseguirán en caso de no llegar a un acuerdo. Para tomar en cuenta este dato pueden hacerse dos cosas:

1) seguir el mismo procedimiento empleado por la función de Nash, e.d., transformar las funciones de utilidad de los jugadores asignando el valor (0,0) al resultado no cooperativo y operar con estas nuevas funciones. Una vez hallada la solución se efectúa la transformación inversa.

2) Hacer un cálculo directo. En este caso, el principio se transformaría del siguiente modo: el jugador 1 tiene que hacer una concesión si $U_1 - U_2/U_1 - U^* < V_2 - V_1/V_2 - V^*$, y el jugador dos tendrá que conceder en caso de que la inecuación aparezca con signo contrario. Como se recordara, esta alternativa también se presentaba en el caso de la función de Nash. Al igual que sucedía entonces, ambos medios de tomar en cuenta las utilidades de partida son equivalentes. Por comodidad de cal-

culo, emplearemos por lo general el segundo método.

Al emplear este método, la equivalencia del principio de concesión Zeuthen-Harsanyi con la solución de Nash se muestra en que la propuesta (U_n, V_n) que hace $C_1 = C_2 = 0$ es aquella para la que se cumple $(U_n - U^*) (V_n - V^*) = (U - U^*) (V - V^*)$ para todo (U, V) que sea una posible solución al problema de regateo.

APENDICE 7

La solución al problema de regateo planteada por Gauthier es a grandes rasgos la siguiente.

Al inicio del proceso de regateo, cada uno de los jugadores propone un término de acuerdo lo mejor posible para él. A partir de esas propuestas de los jugadores, generalmente incompatibles, se inicia un proceso en el cual cada uno de los jugadores debe hacer concesiones. El acuerdo cooperativo consistirá en un resultado que haga iguales las concesiones de los jugadores.

Hasta aquí no hay nada nuevo. La diferencia con el procedimiento H-Z-H reside en la medida de la concesión que supone para los jugadores la aceptación de un determinado resultado. Gauthier define dos medidas de concesión:

-concesión absoluta, cuya magnitud estaría definida por la diferencia entre el resultado inicial pretendido por el jugador y el resultado propuesto como término del acuerdo

-concesión relativa, que se expresa como el cociente entre la concesión absoluta y la concesión completa, que a su vez se entiende como la diferencia entre el resultado inicial pretendido y el resultado que conseguiría si no se llegara a un acuerdo.

Es decir, sea U^* la utilidad que un jugador conseguiría en caso de no llegar a un acuerdo o, por emplear la terminología de Gauthier, la utilidad que un jugador tienen en la posición inicial de regateo, U_+ la utilidad pretendida como *desideratum* por un jugador y U la utilidad que le ofrece una determinada propuesta. Entonces, la medida de su concesión absoluta será $U_+ - U$, la de su concesión completa $U_+ - U^*$ y la de su concesión relativa $U_+ - U / U_+ - U^*$

A partir de la definición de la concesión relativa, y utilizando el principio de concesión de Zeuthen, Gauthier formula el *principio minimax de concesión relativa*, según el cual un resultado será seleccionado como término del acuerdo si y sólo si la mayor concesión relativa que exige es la menor posible.

Este principio equivale a exigir que las concesiones relativas que realizan los jugadores sean iguales, e.d., que se seleccione el resultado que conlleve la menor concesión igual para los jugadores. Sea C_i la

concesión relativa del jugador 1 y C_2 la concesión relativa del jugador 2. Si $C_1 \neq C_2$ entonces o bien $C_1 > C_2$ o bien $C_1 < C_2$. Ahora bien, si hay un resultado posible que haga $C_1 = C_2$ y ocurre que $C_1 > C_2$, entonces existe un resultado que hace menor a C_1 , y si ocurre que $C_1 < C_2$, entonces existe un resultado donde C_2 es menor, siendo este resultado en cualquiera de los dos casos el resultado que hace ambas concesiones iguales. Por lo tanto, si el resultado que hace las concesiones iguales es un resultado posible, entonces será el resultado seleccionado por la aplicación del principio¹.

Puede observarse que el procedimiento de Gauthier es similar en muchos aspectos al procedimiento N-Z-H. En primer lugar, ambos procedimientos seleccionan el resultado que hace que las concesiones de los jugadores sean iguales. Y, en segundo lugar, el cociente que mide la concesión relativa de los jugadores es similar en ambos procedimientos. Sin embargo, es en la medida de la concesión donde Gauthier se separa del procedimiento N-Z-H.

La diferencia sería la siguiente. El principio de concesión de Zeuthen-Harsanyi establece que el jugador 1 debe hacer una concesión si

$$U_1 - U_2 / U_1 - U^* < V_2 - V_1 / V_2 - V^*$$

y que el jugador 2 debe hacer una concesión en el caso de que la inecuación cambie de signo. Por otro lado, el procedimiento de Gauthier exige que se proponga un resultado más favorable a 1 si

$$U_+ - U / U_+ - U^* < V_+ - V / V_+ - V^*$$

y que el resultado sea más favorable a 2 en caso contrario.

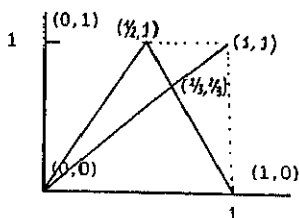
Mientras que en la inecuación de Zeuthen-Harsanyi los únicos valores constantes son U^* y V^* , en la inecuación de Gauthier aparecen otros dos valores constantes, U_+ y V_+ . Es decir, mientras que el principio de concesión Zeuthen-Harsanyi cada jugador mide el valor de la concesión que

¹ Gauthier hace notar que el principio minímax de concesión relativa requiere la menor concesión igual posible sólo en los casos en los que i) el resultado que hace $C_1 = C_2$ es un resultado posible y ii) ese resultado es óptimo. Es decir, si existe un resultado con estas dos propiedades, entonces cualquier otro resultado dará a uno de los jugadores menor utilidad y hará por tanto que su concesión sea mayor. Por tanto, si es posible $C_1 = C_2$, un resultado que exija C'_1 distinta a C'_2 sólo será aceptable si $C'_1 < C_1$ y $C'_2 < C_2$ y la mayor de las concesiones es la menor posible. Ahora bien, como el resultado que se busca a través del regateo es un resultado óptimo, basta con decir que si es posible un resultado que haga iguales las concesiones, entonces ese resultado será seleccionado por el principio. Podemos decir por tanto que las concesiones serán iguales a menos que alguien puede beneficiarse de la desigualdad sin que esto suponga un coste para el otro.

haría de aceptar la propuesta del otro jugador tomando como punto de referencia la utilidad que le proporciona el último resultado que el mismo ha propuesto, el método de Gauthier mide la concesión tomando como punto de referencia la utilidad que le proporciona el resultado más ventajoso para él, esto es, el resultado que el ha propuesto en un principio.

A pesar de esta diferencia, ambos métodos conducen al mismo resultado en casi todos los casos. Sin embargo, Gauthier plantea un caso en el que ambos procedimientos conducen a resultados distintos.

La representación gráfica de este caso es la siguiente



Los resultados posibles están representados por los puntos $(0,0)$, $(1/2,1)$ y $(1,0)$ y los puntos óptimos son los que caen en la línea que une $(1/2,1)$ y $(1,0)$.

Supongamos que los dos jugadores inician un proceso de regateo para elegir uno de estos puntos óptimos. Cada uno de ellos comienza proponiendo el mejor resultado posible para él, esto es, el jugador uno propondrá el punto $(1,0)$ y el jugador 2 el punto $(1/2,1)$. Si aplicamos el procedimiento N-Z-H veremos que el punto resultante será $(1/2,1)$, mientras que si aplicamos el principio de Gauthier el resultado será $(2/3, 2/3)$.

Si comparamos este caso con otros en los que ambos procedimientos conducen al mismo resultado (ver Apendice 6) observaremos algunas diferencias relevantes. En primer lugar, mientras que en los casos anteriores la suma de lo obtenido por ambos jugadores era la misma en todos los resultados posibles, en este caso esto no sucede. En segundo lugar, en este caso las situaciones en las que se encuentran los jugadores no son simétricas y esta asimetría está producida por una característica del juego que no está presente en los casos anteriores. En efecto, en los casos que hemos analizado hasta ahora, la única asimetría que aparecía estaba relacionada con los distintos resultados que los jugadores conse-

guían en caso de no cooperación. Es decir, era una asimetría representada en el *status quo*. Esto no sucede aquí. Por el contrario, en este caso la posición inicial de regateo es igual para ambos jugadores, a saber, el punto (0,0). La asimetría aquí viene dada por el hecho de que no se cumple que si (U,V) es un resultado posible, entonces también lo es (V,U). Es decir, aunque (1/2,1) es un resultado posible, no lo es (1,1/2) etc.

Este ejemplo es difícil de interpretar. Podemos empezar por preguntarnos a qué obedecen las propuestas iniciales de los jugadores. Según Gauthier, cada uno de los participantes en el proceso de regateo empieza por exponer su pretensión máxima, e.d., lo que idealmente desearía ganar con la cooperación. Concretamente:

"cada persona espera que sus logros estarán relacionados con sus pretensiones. Cada uno quiere conseguir tanto como pueda; por lo tanto cada uno pretenderá tanto como le sea posible. Pero, a la hora de decidir qué es lo posible, cada uno se ve constreñido por el reconocimiento de que ni debe hacer que los otros abandonen la mesa de negociación ni tampoco verse excluido por ellos. Por lo tanto, la pretensión de cada persona está limitada por la plusvalía general de la cooperación, y más concretamente, por la porción de esta que le es posible a él recibir. Pretender más sería proponer que los otros renunciaran a algo de lo que traían a la mesa de negociación, a parte de sus ganancias en la posición original de regateo. Puesto que ningún agente racional puede esperar que otro agente racional haga eso, pretender más ... sería ocioso, o, lo que es peor, si uno insistiera en tal pretensión desmedida lo único que conseguiría sería que, o bien los demás abandonaran la negociación, o bien que le excluyeran a él de la misma"².

Gauthier en este punto tampoco parece separarse de la teoría anterior, pues lo que dice en este texto es que la pretensión máxima de cada agente, a partir de las cuales se desarrolla el regateo, consiste en pretender para sí mismo todo el beneficio posible dejando a los demás con la misma ganancia que obtendrían en caso de no cooperación. Esto parece sumamente sensato, pues difícilmente puede esperarse que alguien acepte cooperar en unos términos que le dejarán peor parado de lo que saldría si no cooperara.

Volviendo a nuestro caso, el jugador uno pretende un resultado (1,0) y el jugador 2 el resultado (1/2,1). Al mismo tiempo, el ejemplo

² (Gauthier, 86 pp.133-4)

supone que la posición inicial de regateo es el punto $(0,0)$. Ahora bien, si esto es así, entonces tenemos que suponer que el jugador 2 no está pretendiendo todo lo que podría. Porque, si en la posición original de regateo el jugador uno obtiene una ganancia de 0 ¿porque 2 le ofrece de entrada 0.5? Si suponemos que ambos jugadores pretenden sacar el mayor partido posible de la cooperación y si además suponemos que lo que cada uno consiga estará en relación con sus pretensiones iniciales, entonces lo lógico es suponer que 2 pretenderá inicialmente un resultado que le de a él toda la ganancia de la cooperación, en nuestro caso $(0,1.5)$. Sin embargo, como puede observarse en la representación gráfica, este resultado no es posible. En el caso presentado por Gauthier, los valores que representa la utilidad del jugador 1 en los distintos resultados óptimos posibles, e.d., los valores de U , están comprendidos entre $1/2$ y 1, mientras que los del jugador 2, e.d., los valores de V se encuentran entre 0 y 1.

Podría pensarse que lo que esto quiere decir es que lo peor que le puede pasar a 1 es conseguir $1/2$ y lo peor que le puede pasar a 2 es conseguir 0. Pero, si esto fuera así, entonces no se entendería bien porque Gauthier insiste en que la posición inicial de regateo es $(0,0)$. Más bien, si en el peor de los casos 1 consigue $1/2$, la posición inicial de regateo debería ser $(1/2,0)$. En realidad, lo que parece ocurrir es lo siguiente. Si los jugadores no llegan a un acuerdo y el juego se desarrolla de forma no cooperativa, entonces el resultado será $(0,0)$. Si cooperan, entonces los puntos óptimos aparecen en la línea que une $(1/2,1)$ y $(1,0)$, e.d., son los puntos definidos por la ecuación $Y = (1 - X) / 0.5$. Esto quiere decir que una vez que los jugadores deciden cooperar lo menos que conseguirá 1 es $1/2$ y lo menos que conseguirá 2 será 0.

Podemos ahora preguntarnos qué situación posible está representada en este juego. Para ello, debemos empezar por explicar cómo puede suceder que la suma de lo obtenido por ambos jugadores no sea la misma en todos los resultados posibles. Supongamos que se trata de un problema de distribución. Los dos jugadores tienen que repartir algo, pongamos 100 pesetas. Por el motivo que sea, si no se ponen de acuerdo no tendrán nada que repartir. Dos de los posibles acuerdos son $(100,0)$, $(0,100)$. Supongamos que las utilidades se asignan de la forma habitual, e.d., 0 para la peor alternativa y 1 para la mejor. Por tanto, las utilidades de estos

resultados serían $(1,0)$, $(0,1)$. En el ejemplo de Gauthier ocurren dos cosas extraordinarias, a saber, que no es posible $(0,1)$ y que es posible $(1/2,1)$. Es decir, no es posible el resultado $(0,100)$ y es posible el resultado $(50,100)$!. Naturalmente, esto no puede ocurrir. De modo que este caso no es un caso de distribución, al menos si suponemos que la utilidad es lineal con el dinero.

Podemos eliminar este último supuesto y entender que las utilidades obedecen a otros factores. Por ejemplo, podemos suponer que para el jugador uno el sólo hecho de llegar a un acuerdo le proporciona una utilidad de $1/2$, ya que valora extraordinariamente la cooperación. Esto explicaría el resultado $(1/2,1)$ como correspondiente a la distribución $(0,100)$. El problema está ahora en cómo interpretar la valoración de otros resultados, por ejemplo el resultado seleccionada por Gauthier como el punto de acuerdo, $(2/3,2/3)$. Si la utilidad de 2 depende sólo del dinero conseguido y es lineal con este, debemos interpretar que 2 consigue alrededor de 66 pesetas. Es de suponer que entonces 1 consigue aproximadamente 34. Si su utilidad se incrementa linealmente con el dinero, entonces 1 conseguirá una utilidad de $1/3$ por el dinero conseguido, y otro $1/3$ por el valor asignado a la cooperación. Esto sin embargo es improbable, pues el valor de la cooperación para 1 es $1/2$. Suponer que valora más la cooperación cuando no consigue nada que cuando consigue algo es improbable. Podemos suponer que 1 es rico, de modo que la utilidad que tienen para el 34 pesetas es sólo de 0.16. Esto explicaría el resultado $(2/3,2/3)$. También explicaría otros resultados, por ejemplo, el resultado $(1,0)$. Aquí, la utilidad de 1 resultaría del $1/2$ que obtiene del puro hecho de cooperar más la porción correspondiente de la utilidad que le dan 100 pesetas, que es aproximadamente del valor que la misma cantidad tiene para el otro, o sea, $1/2$.

Aunque esta explicación es posible, no deja de resultar extraordinaria. El jugador 1 valora la cooperación al margen de lo que el otro consiga. Lo mejor para él es llevarse todo, e.d., las 100 pesetas, con el consentimiento del otro (dejando a parte que por la definición del juego tal consentimiento es necesario). Pero ser extraordinario no es por sí una objeción. Lo que sí que es una objeción es que esta hipótesis no asigna utilidades de la manera habitual. Es posible que el dinero valga menos para uno que para otro y que el hecho de cooperar también se valore

de forma distinta para cada uno de los jugadores. Pero esto cae fuera de la cuestión. Para empezar, hacer este tipo de juicios requeriría algún tipo de mecanismo para realizar comparaciones interpersonales de utilidad, cosa que Gauthier rechaza explícitamente. Cada uno puede hacer sus valoraciones como quiera, pero se supone que la asignación de utilidades se realiza de tal forma que cada uno de los jugadores asigna el valor convencional 0 a 1 menos preferida y 1 a la más. Estos dos valores límite pueden ser los que se quiera, pero deben ser los mismos para ambos jugadores. De modo que la alternativa (0,100), supuesto que es la menos preferida por 1, puede valer en realidad para 1 más, mucho más, menos mucho menos, igual, más o menos igual, etc. que lo que vale para 2 (100,0). Pero el juego debe realizarse asignando los mismos valores límite.

Como, por otro lado, Gauthier utiliza este método de asignar utilidades y rechaza la comparación interpersonal, tenemos que suponer que la explicación que hemos dado del ejemplo no es la correcta. Podemos intentar dar otra. Si el ejemplo no puede ser un caso de problema de distribución, podemos suponer que se trata de un problema de producción. Los jugadores quieren ponerse de acuerdo para producir unos beneficios que luego se repartirán. Desde luego, el regateo también se establece para llegar a un acuerdo acerca de como se repartirán lo producido. Pero, además, en este caso es de suponer que habrá por ambas partes, o al menos por alguna de ellas, costos de producción. Es decir, para los jugadores la cooperación supondrá algún tipo de coste, bien sea en dinero, trabajo, talento o lo que se quiera.

Esta interpretación tampoco resulta satisfactoria. Gauthier supone que cada uno puede esperar beneficios de la cooperación en proporción a su contribución a la misma. Esto puede representarse de dos modos

- reflejando la contribución en la posición inicial de regateo, y jugando un juego de regateo en el que se trata de distribuir lo que podemos llamar beneficios brutos

- haciendo que la posición inicial de regateo sea (0,0) y jugando un juego de regateo para distribuir beneficios netos, e.d., entendiéndolo que los costes de los jugadores ya se han cubierto.

Veamos que pasa si interpretamos el ejemplo en el que el método de Gauthier y el procedimiento N-Z-H difieren como un caso de este tipo. Hay tres posibilidades:

1- que los costes ya estén cubiertos y los distintos resultados posibles reflejen distintas distribuciones de beneficios netos. Cada uno propone como desideratum quedarse con la totalidad del beneficio neto. El jugador 1 propone (1,0) y el jugador 2 propone (1/2,1). de nuevo, vuelven a surgir los mismos problemas de antes: no se explica porque 2 le ofrece de entrada 1/2 al jugador 1 y no se entiende como es posible que el beneficio neto cambie de valor, e.d., que valga 1 en un caso y 1.5 en otro.

2- que los costes estén incluidos en los pagos, e.d., que los pagos representen beneficios brutos. Cada uno propone que el otro salga de la aventura cooperativa con los costes cubiertos mientras que el se lleva todo lo demás, a saber, la parte con lo que cubre sus propios costes más el beneficio neto. El jugador uno propone (1,0) por que el jugador 2 no ha tenido costes de producción. El jugador 2 propone (1/2,1) porque los costes de 1 son de 1/2. Esto tampoco es posible. Si los costes de 1 son 1/2, entonces el beneficio neto es de 1/2, de modo que 1 sólo puede conseguir 1 si no 2 no cubre costes. Igual que el beneficio neto no puede ser a veces de un valor y otras de otro, lo mismo le ocurre al beneficio bruto.

3- que los costes del contrario estén cubiertos, pero no así los del jugador que hace la propuesta. Es decir, cuando los costes de 2 se cubren queda 1 para repartir. Naturalmente, 1 propondrá quedarse con 1 con lo que cubriría sus propios costes y se quedaría con todo el beneficio neto. El jugador 2 propone (1/2,1). De nuevo, ¿porque le ofrece 1/2 al contrario si los costes del contrario ya están cubiertos? Aparte de esto, tampoco es posible explicar los pagos. Por lo siguiente. Sean C_1 y C_2 los costes respectivos de los jugadores 1 y 2, BN_1 y BN_2 los beneficios netos, BB_1 y BB_2 los beneficios brutos, TB el total bruto y TN el total neto.

A partir de la propuesta (1,0) podemos inferir que

$$TB = BB_1 + BB_2 = BN_1 + C_1 + 1 = C_1 + 1$$

mientras que a partir de la propuesta (1/2,1) sabemos que

$$TB = BB_1 + BB_2 = 1 + BN_1 + C_1 = 1 + 1/2 + C_1 = 3/2 + C_1$$

Si el total bruto es de 1 una vez que se descuentan los costes del jugador 2 y de 3/2 al descontar los costes del jugador 1, entonces $C_1 = C_2 + 1/2$. Por tanto, $TB = TN + C_1 + 1/2 + C_2 = TN + 2C_1 + 1/2$.

A partir de estos datos podemos interpretar el ejemplo de Gauthier

del siguiente modo:

$$C_1 = 0, C_2 = 1/2$$

$$TN = 1, TB = 3/2$$

Esto explicaría la propuesta del jugador 1, (1,0). Este jugador descuenta del total bruto los costes del jugador 2 y propone quedarse con el resto. Pero, ¿cómo podríamos explicar la propuesta del jugador 2? Puesto que los costes del contrario son 0, debería proponer inicialmente (0,3/2). Supongamos que de hecho hiciera esa propuesta, que es la que la teoría exige. Como ya dijimos, esto es equivalente a un juego en el que la posición inicial de regateo sea (0,1/2) y las propuestas iniciales (1,1/2) y (0,3/2). Si jugamos un juego de regateo con estas características, entonces, tanto aplicando el principio *N-Z-H* como aplicando el principio de Gauthier, el resultado del regateo es (1/2,1). El resultado (1/2,1) equivale a repartir a partes iguales el beneficio neto. La diferencia entre el pago de uno y otro jugador obedece a que el jugador uno sólo recibe su porción correspondiente del beneficio neto (1/2) mientras que el jugador dos recibe una parte igual de este beneficio más el pago de sus costes (1).

Por tanto, en la única interpretación que se me ocurre del caso presentado por Gauthier, la propuesta inicial del jugador 2 no es la adecuada. Si lo fuera, entonces el resultado sería (1/2,1). Hemos mostrado que el caso presentado por Gauthier tiene un resultado igual al aplicarse ambos principios si se resuelve empleando el juego equivalente en el que la posición inicial de regateo es (0,1/2) y las propuestas iniciales (1,1/2) y (0,3/2). Podemos preguntarnos que pasaría si se jugara de la forma originaria, a saber, con un posición inicial de regateo de (0,0) y unas propuestas iniciales de (1,0) y (0,3/2). Como dijimos, la diferencia de la suma de las utilidades conseguidas por los jugadores se explica por el hecho de que estos pagos suponen cubiertos los costes del contrario. Esta forma de plantear el juego es bastante confusa, pues en cada par (U,V) hay que entender que, si la propuesta es del jugador 1, entonces U expresa beneficios brutos y V beneficios netos y al contrario si la propuesta es del jugador 2, pero Gauthier la emplea en ocasiones (por ejemplo, ver Gauthier 86, pp.137-9).

Por otro lado, si se efectúa el juego de regateo según esta variante, que intuitivamente es equivalente a la anterior, el resultado no

sería $(1/2, 1)$. Empleando el procedimiento de Gauthier el resultado sería algo más favorable a 1 a costa de 2, concretamente $(0.6, 0.9)$, lo cual es intuitivamente incorrecto, por que si al valor de V le descontamos los costos y se expresa el resultado como beneficios netos tendríamos un pago de $(0.6, 0.4)$ lo cual no parece un reparto muy justo. Y El procedimiento N-Z-H ni siquiera podría emplearse sensatamente. La única posibilidad esta en plantear el juego con unas propuestas iniciales de $(1, 0)$ y $(0, 3/2)$ pero partiendo de un status quo de $(0, 1/2)$. Con esto, el resultado por ambos procedimientos expresado en beneficios brutos vuelve a ser $(1/2, 1)$.

Resumiendo, el único modo de interpretar el ejemplo de Gauthier es suponer que los pares (U, V) , que expresan las peticiones iniciales de los jugadores, deben entenderse del siguiente modo: si la propuesta es del jugador 1, U expresa beneficios brutos y V beneficios netos y al contrario si la propuesta es del jugador 2. A la vista de las peticiones iniciales sabemos que la diferencia de los costes de los jugadores es de $1/2$ más en el caso del jugador 2 que en el caso del jugador 1. Una posible asignación de valores que explicaría el par $(1/2, 1)$ es que el coste del jugador 1 sea 0, el del jugador 2 de $1/2$ y el beneficio total neto de 1. Tomando en cuenta estos valores, si ambos jugadores pretenden inicialmente el total del beneficio neto, dejando al contrario con los costes cubiertos, el procedimiento N-Z-H y el procedimiento de Gauthier conducen al mismo resultado $(1/2, 1)$ en cualquier modalidad equivalente del juego.

Sin embargo, Gauthier tiene razón al afirmar que el procedimiento N-Z-H y el suyo propio conducen en ocasiones a resultados distintos. Por ejemplo, esto sucede en el caso presentado por Gauthier si admitimos como validas las propuestas iniciales de los jugadores. El jugador 1 propone $(1, 0)$, e.d., propone quedarse con todo el beneficio neto suponiendo cubiertos los costes de 2. Si el jugador 2 hiciera una propuesta similar debería proponer, como ya hemos dicho, $(0, 3/2)$. Pero supongamos que, por el, motivo que sea, el jugador 2 hace un ofrecimiento inicial de $(1/2, 1)$. Por lo tanto, el regateo se desarrolla a partir de estas propuesta y con

una posición inicial de regateo de $(0,1/2)$ ³. ¿Que sucede en este caso? Aplicando el procedimiento N-Z-H el resultado sigue siendo $(1/2,1)$ entendiendo ambos valores como beneficios brutos. Puesto que los costes de los jugadores son $(0,1/2)$, este resultado equivale a repartir equitativamente el beneficio neto. Es decir, el resultado es el que intuitivamente debe ser.

Si utilizamos el procedimiento de Gauthier el resultado aproximado es $(0.67,0.83)$, también en valores brutos. Este quiere decir que si le descontamos al jugador dos sus costes, su beneficio neto es de 0.33, frente al beneficio neto de 1 que es de 0.67. Este resultado no es muy intuitivo. Si el jugador 2 es un jugador racional, ¿por que iba a aceptar un resultado tan desfavorable para él? Una posible respuesta de Gauthier sería: Si 2 es un jugador racional, ¿porque empieza proponiendo $(1/2,1)$ en vez de $(0.3/2)$?. si hubiera empezado con esta última propuesta, le hubiera ido mejor. Como dice Gauthier en el texto citado (p.321), el resultado dependerá de las propuestas iniciales, por lo cual cada uno estará interesado en proponer de entrada el resultado que le sea más favorable. Precisamente en este punto reside la diferencia entre ambos procedimientos. Citando al propio Gauthier "mientras que en nuestro procedimiento cada uno tiene interés en expresar su pretensión máxima, en este procedimiento alternativo [el procedimiento N-Z-H] nadie tiene ese interés, puesto que nadie puede esperar conseguir más pretendiendo más" (86, p.148). En efecto, según el procedimiento N-Z-H, si el jugador dos propone $(1/2,1)$, el jugador 2 no puede esperar más mediante al proceso de regateo, pues el contrario ya le ha ofrecido todo lo que puede ofrecerle, ya le ha ofrecido todo lo que 1 puede esperar conseguir de un contrario racional, a saber, un reparto equitativo de los beneficios netos.

Gauthier encuentra que esta característica del procedimiento N-Z-H es muy objetable, especialmente porque no da lugar a una medida sig-

³ Debe tenerse en cuenta que este no es el juego desarrollado por Gauthier en la página 147 de morals by Agreement. El juego presentado por Gauthier se diferencia de este en la posición inicial de regateo, que en el ejemplo de Gauthier es de $(0,0)$. Gauthier llega a un resultado de $(2/3,2/3)$. Pero, si el juego se interpreta como nosotros lo hemos hecho, e.d., entendiendo que cada jugador hace una propuesta suponiendo cubiertos los costes del contrario, entonces el coste no es el mismo para los dos jugadores y esta diferencia debe reflejarse en el origen del juego. De hecho, si no se hace así ¿cómo hay que interpretar el resultado $(2/3,2/3)$ ¿son beneficios netos, brutos, uno neto y otro bruto? Ninguna de estas posibles interpretaciones explica el resultado.

nificativa de las concesiones relativas de los jugadores. Si la concesión se mide a partir de unas propuestas *que no tienen valor real, e.d., que no tienen capacidad de modificar el resultado del regateo*, entonces, ¿qué sentido tiene el valor de la concesión?. Sin embargo, como afirma el propio Gauthier (p.149), podemos suponer que los jugadores deben empezar por exigir lo máximo posible, no porque esto vaya a alterar el resultado del regateo, sino porque de ese modo obtenemos un valor de las concesiones relativas con significado. En el procedimiento N-Z-H, cada uno debería empezar exigiendo lo máximo para hacer ver al contrario lo que va a conceder.

Aunque el procedimiento N-Z-H pueda dar significado a la magnitud relativa de las concesiones de los jugadores si estos empiezan el juego formulando su pretensión máxima, sigue siendo cierto que, en los casos en que alguno de los jugadores empieza con una pretensión menor a la máxima, ambos procedimientos conducen a resultados distintos. Esto quiere decir que si el procedimiento de Gauthier selecciona el resultado que exige la mínima concesión posible por parte de los jugadores, entonces en los casos en los que el resultado seleccionado por el procedimiento N-Z-H es distinto sucederá que este resultado exige una concesión mayor que la mínima posibles. Ahora bien, parece que es exigible como condición de la racionalidad de un acuerdo el que no requiera una concesión mayor que la mínima posible. No hay ningún motivo por el que un jugador racional debe acceder a un acuerdo que le exija conceder más, ya que esto supone que el acuerdo tendrá para el una utilidad menor de la posible. Por consiguiente, Gauthier encuentra que toda alternativa a su método, e.d., toda alternativa que seleccione un resultado distinto al seleccionado por el principio minimax de concesión relativa, es por ello mismo criticable.

Sin embargo, como ya dijimos (pp.77 y ss.) puede mostrarse que la solución de Nash es la única que cumple con unas condiciones que parecen intuitivamente exigibles (el principio de concesión Zeuthen-Harsanyi cumple con estas condiciones en tanto que es un método equivalente al de Nash). Puesto que el procedimiento de Gauthier es distinto, podemos suponer que incumple con alguna de aquellas cuatro condiciones. En efecto, la crítica de Gauthier al procedimiento N-Z-H pasa por la crítica de una de estas condiciones que, según argumenta Gauthier, es la responsable de que dicho procedimiento seleccione un resultado determinado con independencia

de las peticiones con las que los jugadores inicien el regateo : la independencia de las alternativas irrelevantes.

Como se recordara, esta condición dice que, dado un juego de regateo G , si restringimos el conjunto de los resultados alternativos dando lugar a un nuevo juego G^* sin alterar por ello el status quo, la solución (U,V) de G^* será también la solución de G si (U,V) sigue siendo uno de los pares alternativos de G^* .

En nuestro ejemplo, esta condición se aplica del siguiente modo:

Sea G un juego de regateo cuyo status quo es el par $(0,1/2)$ y cuyo conjunto de resultados óptimos alternativos cae en la línea que une los puntos $(0,3/2)$ y $(1,0)$. El resultado de este juego es el punto $(1/2,1)$. Sea G^* una versión restringida de G , en el cual las alternativas se encuentran en la línea que une los puntos $(1/2,1)$ y $(1,0)$. Puesto que la solución de G es un punto posible de G^* , la solución de G^* será también $(1/2,1)$.

El procedimiento N-Z-H cumple esta condición, pero no así el procedimiento propuesto por Gauthier, que si bien selecciona el punto $(1/2,1)$ como solución a G , selecciona un punto distinto para G^* . Por tanto, parece que la cuestión esta en si es exigible o no esta condición. Sin embargo, hay que tener en cuenta que Harsanyi deriva su solución al problema de regateo a partir de un conjunto de postulados entre los que no se encuentra la independencia de las alternativas irrelevantes (Harsanyi, 77). Harsanyi exige que el resultado del proceso de regateo sólo debe depender de aquellos factores que entran en la elección entre la propia propuesta y la del oponente. Gauthier concede que esto se sigue de la visión del proceso de regateo ofrecida por Harsanyi, según la cual el proceso de regateo consiste en una serie de propuestas sucesivas de los jugadores que son evaluadas y rechazadas hasta llegar a un punto de acuerdo. Pero para Gauthier, si el resultado del regateo es independiente de las pretensiones de los jugadores, entonces todo el proceso de regateo "se convierte desde el punto de vista racional en una farsa".

¿Es realmente una farsa el proceso de regateo?. Para Gauthier lo es desde el momento en que el resultado de este proceso es independiente de lo que inicialmente pretendían conseguir los jugadores. Sin embargo, es necesario hacer algunas matizaciones. Supongamos que tu y yo vamos a repartirnos 100 pesetas. Yo propongo que cada uno nos llevemos 50, y tu

propones quedarte con las 100 y dejarme a mí sin nada. En cierto sentido, el proceso de regateo se convierte en una farsa, a saber, en el sentido de que el resultado del regateo va a ser el que yo he propuesto. Pero esto no significa que yo no cedo nada ya que el resultado va a ser el que yo quiero. Significa que con mi propuesta *yo ya he cedido todo lo que puedo ceder*. Si soy racional no voy a ceder más y, lo que es más importante, *tu lo sabes*, al igual que yo no he pedido más *porque se que tu no vas a ceder más*. Desde luego que es inútil seguir discutiendo. Puede decirse si se quiere que en este caso el regateo no tiene sentido, pero esto no es una objeción. Cuando vamos a comprar algo que vale 1.000 pesetas pueden suceder dos cosas: que el vendedor pida 1.500 o que pida el precio justo. Si hace lo primero, lo hace suponiendo que vamos a regatear. Esto tiene algunas ventajas, ya que la gente, tras regatear, suele irse convencida de que ha conseguido una ganga. Pero esta impresión no deja de ser eso, una impresión: nadie va a vendernos nada por menos de lo que vale (o, si queremos, por menos de lo que pueda obtener). De modo que es de esperar que sea inútil regatear con el vendedor que pide por el artículo lo que este vale. En estos casos el regateo es inútil *porque es innecesario*. El objeto del proceso de regateo es conseguir un acuerdo lo más favorable posible para uno mismo, de modo que si el contrario ya nos ofrece lo más favorable posible no podemos hacer otra cosa más que acceder.

La cuestión está en que Gauthier pone un énfasis excesivo en las propuestas iniciales de los jugadores. Lo realmente importante no es lo que cada uno de los jugadores pretende sino lo que podría pretender. El cociente que utiliza el procedimiento de Zeuthen y Harsanyi mide la ventaja relativa que para un jugador tiene su propia propuesta comparada con la propuesta del contrario y, en consecuencia, la determinación con la que un jugador defenderá su propia propuesta y se negará a aceptar la de su oponente. Si la propuesta que yo te hago es mejor para ti que la que tu haces lo es para mí, yo estaré más decidido a defender mi propuesta que tu.

Los dos procedimientos conducen a los mismos resultados cuando los jugadores empiezan proponiendo su pretensión máxima. La diferencia reside en que con el procedimiento de Gauthier el resultado depende de esta pretensión. Gauthier afirma que, si bien él no cree que la intuición tenga

que decir la última palabra decisiva en cuestiones concernientes a la teoría de la decisión, si tiene algo que decir no es precisamente a favor del procedimiento N-Z-H. Lo poco intuitivo de este procedimiento esta en que, en el caso que hemos discutido, el resultado es el que el jugador 2 quería y, por tanto, toda la concesión es la que hace el jugador 1. Desde luego, presentado de ese modo el resultado no parece muy intuitivo. Pero es dudoso que esa descripción de lo que sucede en el juego sea la adecuada. Parece mucho más ajustado a la realidad decir que la pretensión de 2 es tan "modesta", que ha cedido ya tanto al hacerla que difícilmente el otro jugador puede esperar que conceda más. El jugador 2 ha propuesto ya lo mínimo que esta dispuesto a aceptar y, por ello, su determinación a mantener su propuesta es muy grande. Planteado así, parece que la intuición se inclina más a favor de este procedimiento que en su contra.

APENDICE 8

La idea general de la propuesta de Gauthier es que en el punto de partida de la negociación debe reflejar lo que cada uno de los participantes podría obtener por sus propios medios y sin aprovecharse del resto de los participantes. Esta idea tiene su antecedente en la teoría de la propiedad de Locke (63) y supone una extensión, así como una modificación, de esta teoría.

Según Locke, cada uno tiene derecho a los frutos de su propio trabajo, siempre y cuando se cumplan dos condiciones, a saber, que se use, o al menos no se malgaste, lo producido y que se deje para los demás una cantidad "suficiente e igualmente buena" de lo que se toma. Gauthier apunta que las palabras de Locke tomadas literalmente son inaplicables. El motivo es que no podría haber apropiación de ningún bien escaso o potencialmente escaso, pues cuando se trata de bienes de este tipo, puesto que no puede haber suficiente de lo que es escaso, si uno toma suficiente para sí no puede dejar suficiente para los demás. Pensemos por ejemplo en el agua. En un sitio donde el agua sea abundante yo puedo tomar tanta agua como quiera y tengo derecho a hacerlo siempre y cuando me la procure yo mismo. Puede beber tanto como quiera, regar mis tierras hasta convertirlas en un vergel y ducharme cuarenta veces al día si ese es mi gusto. Aun así, habrá para los demás tanta agua como la que yo misma tomo. Pero en un sitio donde el agua es un bien escaso la cosa cambia. Si yo utilizo todo el agua que quiero entonces dejo a los demás con menos agua de la que yo mismo tomo. Este inconveniente es serio, y no sólo porque hay muchos bienes escasos, sino por algo mucho más fundamental. Sólo con los bienes escasos surge el conflicto de intereses y, por tanto, la necesidad y la posibilidad de negociar. Sólo donde el agua es un bien escaso hay tribunales de agua porque sólo allí son necesarios. Donde el agua abunda a nadie se le ocurre discutir por la cantidad de agua a la que tiene derecho¹.

Por este motivo, Gauthier acude a la interpretación de las pala-

¹Gauthier no hace mención especial de este punto, pero es tan evidente que supondremos que no lo menciona por darlo por descontado.

bras de Locke ofrecida por Nozick (74). Para Nozick, la referencia de Locke a "suficiente e igualmente bueno" tiene el propósito de asegurar que, al apropiarse uno de los frutos de su esfuerzo, la situación de los demás no se ve empeorada. Pero como apunta Gauthier, este requisito es demasiado fuerte, ya que hay ocasiones en las que el único modo de no empeorar la situación de los demás es empeorar la propia. Ningún agente racional aceptaría un requisito que exigiera tamaño sacrificio. Por ello, Gauthier propone aceptar la interpretación de Nozick añadiendo la palabra "innecesariamente". Es decir, cada uno tendrá derecho a los frutos de su propio trabajo si con ello no empeorará la situación de los demás, a menos que esto sea necesario para evitar empeorar la situación propia.

Este principio es conocido como *salvaguarda lockeana*, nombre debido a Nozick. La salvaguarda supone una restricción a la maximización de la utilidad individual. En efecto, el principio prohíbe mejorar la propia situación si esto se hace a costa de empeorar la situación de los demás. Empeorar la situación de los demás sólo está permitido en tanto que medio necesario para evitar que la propia situación se vea empeorada. En la formulación del principio aparecen los términos "mejorar" y "empeorar". Estos son términos comparativos. Por tanto es preciso establecer el término de comparación que nos servirá para decidir cuando la situación de los demás empeora o cuando mejora la situación propia.

Con el fin de señalar este punto de comparación, Gauthier utiliza un ejemplo sumamente clarificador. Supongamos que alguien, a quien llamaremos X, está ahogándose en las aguas de un río. En ese momento otra persona, a quien daremos el nombre de Y, oye los gritos de socorro de X. Haciendo caso omiso de esta llamada de socorro, Y sigue su camino dejando que X se ahogue. La cuestión es si Y ha empeorado o no la situación de X. Se da por supuesto que la situación de X es peor cuando se ahoga que cuando estaba vivo. Esto no se discute (se podría discutir en relación a otro problema, pero para el tema que nos ocupa podemos darlo por supuesto). Lo que se discute es si es Y es causante de este empeoramiento de la situación de X.

Para contestar esta pregunta tenemos que saber lo que ha pasado antes del momento en el que hemos tomado la historia. En efecto, pueden haber sucedido dos cosas. Puede ser que X haya tropezado cayendo al río y que Y pasara por allí casualmente. O puede suceder que haya sido Y quien

haya empujado a X haciéndole caer. En el primero de los dos casos Y, desde luego, no mejora la situación de X, pero tampoco puede decirse que la empeore. Y esto por un motivo. Si Y no hubiera pasado por allí el resultado para X hubiera sido el mismo, e.d., se habría ahogado. Por el contrario, en el segundo caso es Y quien ha empeorado la situación de X, porque en ausencia de Y el resultado para X hubiera sido distinto. No se habría ahogado.

Este ejemplo ilustra a la perfección lo que para Gauthier debe ser el punto de comparación que permite hablar de "mejorar" o "empeorar" la situación de otro. Para saber si un agente Y empeora o mejora la situación de otro agente X debemos comparar el resultado con el que se hubiera seguido en ausencia de Y. En el caso de que lo que este en discusión sea el empeoramiento o mejoramiento de la situación propia a costa de otros el punto de comparación es paralelo. En este caso, se trata de saber lo que hubiera sucedido si los demás no hubieran estado presentes. En otras palabras, el punto de comparación es el resultado que habría tenido lugar en ausencia de interacción. Por tanto, debe entenderse que la salvaguarda prohíbe mejorar la propia situación mediante una interacción que empeore la situación de aquellos con los que se interactúa.

Según Gauthier, la aceptación de la salvaguarda proporciona una estructura primaria de derechos y deberes. La salvaguarda garantiza a cada uno derecho exclusivo al ejercicio de las propias capacidades, así como el deber de no interferir en el uso de las capacidades ajenas. Estos derechos y deberes se hacen extensivos a lo conseguido mediante el uso de tales facultades, con la diferencia de que en este caso no puede hablarse de derecho exclusivo, ya que la salvaguarda no resulta violada si alguien utiliza los frutos del trabajo de otra persona compensándola por la pérdida de utilidad sufrida al no disfrutar de lo que le pertenece.

APENDICE 9

Parfit señala cuatro de estos cambios de carácter moral que, de darse en todos los agentes involucrados en un dilema, darían solución a este ¹. En primer lugar, el agente puede convertirse en una persona digna de confianza. Puesto que sabemos que en determinadas situaciones es racional llegar a un acuerdo, los agentes implicados lo harán. Una vez alcanzado este, todos prometerán llevarlo a cabo. Y si son personas dignas de confianza todos cumplirán lo prometido, e.d., todas realizarán la estrategia A.

En segundo lugar, el agente puede hacerse reactivo a ser aprovechado. Un agente con esta característica preferirá cumplir el acuerdo si los demás lo hacen a pesar de que ganaría más no haciéndolo, debido a la repugnancia que le produce obtener beneficios a costa de los demás. Por eso, si sabe que los demás harán A, el también hará A.

En tercer lugar, el agente puede hacerse kantiano. Ante cualquier alternativa, el agente se preguntará si puede querer racionalmente que todos actúen según esa estrategia y sólo la realizará él mismo si la respuesta es positiva. Puesto que la situación en la que todos hacen S es una situación en la que todos salen perdiendo, el agente así dispuesto no realizará B, sino A.

Por último, el agente puede convertirse en una persona altruista. En una situación en la que puede elegir entre una estrategia que S que le beneficia a él y una estrategia A que beneficia al otro, siempre elegirá A.

Cada uno de estos cambios tiene sus inconvenientes a la hora de solucionar el dilema. Sin embargo algunos de estos no son importantes. Por ejemplo, el segundo de estos cambios presenta, en opinión de Parfit, el inconveniente de que, si hay un número suficiente de personas con esta característica, entonces no sería necesaria la existencia de un acuerdo

¹ Es posible que alguien pueda pensar en algún cambio de actitud moral distinto de estos cuatro. Yo no encuentro ninguno, pero, naturalmente esto no quiere decir que no pueda haberlos. Esto sin embargo no me parece de mucha importancia. Mi propósito es analizar el funcionamiento de las soluciones morales en general y, más concretamente, que implicaciones teóricas tiene la existencia de la solución S. Y basta con que esta solución pueda plantearse con cualquier tipo de cambio moral, cosa que parece indiscutible.

condicional². Bastaría con la seguridad de que un número suficiente va a realizar A para que los que tienen esta disposición hicieran A a su vez. Pero los agentes así dispuestos no podría crear esta seguridad. Este inconveniente puede tener su importancia, pero puede ser superado si suponemos la existencia de un acuerdo condicional y suponemos también el conocimiento por parte de cada uno de los agentes de lo que los otros están haciendo.

Otro problema más importante y relacionado con este es quién va a establecer la práctica de la cooperación. Si el único motivo que mueve a los agentes es su repugnancia a ser aprovechados, una vez que existe una práctica en la que todos o un número suficiente cooperan, estos cooperarán también. Pero si esta práctica está aun por establecerse, un agente impulsado sólo por esta motivación no podría dar el primer paso. Si no hay nadie cooperando, no se puede ser un aprovechado por no cooperar. Pongámonos en el puesto de un agente con esta característica que tienen que elegir entre B y A. ¿Cuáles son los motivos para realizar B? Según hemos dicho, el motivo es que si los demás hacen B, lo mejor es hacer B también, y si los demás hacen A también lo es. Esto último es precisamente lo que cambia en este caso. Si los demás o un número suficiente hacen A, entonces lo mejor para mí es hacer A también, porque de otro modo estaría actuando como un aprovechado, cosa que por hipótesis quiero evitar. Por tanto, sólo en el supuesto de que los demás o un número suficiente de los demás van a hacer B, tendré yo motivos para hacer B. Que yo sepa que todos tienen la misma disposición que yo y que todos saben esto de todos los demás no soluciona nada.

El cambio consistente en que los agentes se conviertan en personas dignas de confianza tiene también sus problemas. Si cada uno sólo se comprometerá a realizar su parte si todos los demás también se comprometen, esta característica por parte de los agentes puede solucionar el proble-

² Con "acuerdo condicional" Parfit se refiere al acuerdo de realizar determinada estrategia conjunta si los demás, por su parte, también lo hace. Como ya vimos, la existencia de este acuerdo condicional no garantiza su cumplimiento si el resultado carece de estabilidad. Sin embargo, en la mayor parte de los casos, es condición necesaria para que la estrategia conjunta pueda realizarse.

Es importante no confundir este acuerdo condicional, que en algunos casos parece prescindible, con el acuerdo al que se llega mediante el proceso de regateo acerca de la estrategia conjunta a realizar. Este no es prescindible en ningún caso. Sólo cuando existe este acuerdo es posible plantearse si seguirlo o no, o si seguirlo sólo a condición de que los demás lo hagan también.

ma. Pero otra cosa muy distinta sucede si no se alcanza una unanimidad total. En efecto, puede no ser necesaria ni posible la unanimidad, sino que puede suceder que la cooperación de un número n de personas sea suficiente para lograr el resultado apetecido. Por ejemplo, supongamos que 3 personas quieren ir a jugar al frontón, pero para eso tienen que ponerse de acuerdo en quien irá a por la pista. Como quieren jugar dos horas (menos no vale la pena) tienen que ir dos. Los tres son personas de confianza y si se comprometen a ir, irán. Ahora bien, cada uno puede hacer el siguiente razonamiento. Si los otros dos se comprometen a ir, yo no necesito comprometerme. Y si ninguno de ellos lo hace, entonces no sirve de nada que lo haga yo. Sólo si uno (y sólo uno) de ellos se compromete haré yo bien en comprometerme también. Por supuesto que si me comprometo a colaborar cumpliré mi palabra, pero sólo en el caso de que yo haga el número 2 será razonable que adquiera ese compromiso.

Esta situación es parecida a la que se plantea entre agentes no dignos de confianza. En ese caso vimos que es racional cooperar si mi cooperación es condición necesaria para que el resultado deseado se realice, y si es necesaria la participación de todos los implicados entonces el acuerdo es estable. Pero en otro caso surgía el dilema. Aquí sucede algo similar. La diferencia está en que si los agentes en cuestión son dignos de confianza yo sé que, si se han comprometido a hacer A, harán A. Y si hacen A en un número suficiente, entonces yo no gana nada comprometiéndome a hacer A. Si me comprometo a ello tendré que hacerlo, pero es mejor para mí hacer β . Y si no se han comprometido, también es mejor que haga β .

Sin embargo, y a pesar del parecido, la situación no es la misma. Porque en el primer caso el problema es si yo voy o no voy a cumplir el acuerdo, mientras que en el segundo caso, e.d., en el caso de que se trate de personas dispuestas a mantener el acuerdo, el problema se transforma en qué acuerdo va a lograrse. Puede en efecto efectuarse un regateo para acordar una estrategia conjunta de la forma "X e Y irán y Z no irá". Si nadie puede hacer una presión razonable para ser él el favorecido por el trato, pueden jugárselo a los chinos. Y una vez acordada esta estrategia, todos la seguirán. De hecho, ante una situación como esta, esto es lo que hace la gente digna de confianza. Naturalmente, en situaciones que involucren a más jugadores la situación es más difícil de resolver. Pero

la dificultad *nunca* está en si uno se comprometerá a hacer algo o no. La dificultad está, más bien, en como acordar la estrategia conjunta a seguir. Pero una vez que esta ha sido acordada, cada uno cumplirá su parte. Dicho de otra forma, a lo que cada uno se compromete es a cumplir su parte. Y cual sea esta parte es algo que se decide mediante los procedimientos de regateo indicados.

Puede plantearse en este punto una objeción. Esta surge porque cada uno se compromete a realizar su parte *después* de que se haya acordado que parte corresponde a cada uno. Ahora bien, los acuerdos para los que no se requiere unanimidad son tales que el resultado se producirá aunque no todos cumplan su parte. Hablando entre gente decente, aunque no todos se comprometan a cumplir lo que les ha tocado. En el ejemplo anterior, si se acuerda que vayan los tres, el resultado deseado (la obtención de la pista) se sigue aunque sólo dos se comprometan (y cumplan, cosa que ahora damos por supuesta). Pero, de nuevo, el problema es ¿quién se compromete?. Y el dilema vuelve a plantearse.

Esta objeción, así planteada, puede ser contestada con facilidad. Por definición, el resultado del acuerdo será óptimo. Y un acuerdo que consiste en que vayan tres donde pueden ir dos no lo es. Hay otra situación que es mejor para alguno sin que sea peor para ninguno, a saber, la situación en la que sólo van los necesarios. Por consiguiente, parece que no puede plantearse un acuerdo en el que el costo de uno sea gratuito. Sin embargo, hay otros casos en los que la respuesta no es tan sencilla, a saber, los casos del tipo del dilema del contribuyente. En estos casos la contribución individual se ve como algo insignificante que no hace ninguna diferencia. De modo que si poca gente se compromete a pagar, que yo pague no sirve para alcanzar los bienes públicos deseados, y si pagan los suficientes, seguro que aunque yo no pague no va a pasar nada. De modo que lo mejor es no comprometerme. Por supuesto que si yo no colaboro el resultado no será exactamente igual, pero si infinitamente parecido. Por ello, el que todos seamos dignos de confianza no parece suficiente para solucionar este tipo de casos³.

³ Sin embargo, esta solución es buena si se trata de acuerdos condicionales, donde el compromiso de todos los demás es condición indispensable para que se comprometa cada uno. Podría aun surgir, como en el caso anterior, el problema de quién se comprometa primero, pero esto puede solucionarse de diversos modos, por ejemplo mediante compromisos simultáneos.

El tercer cambio, consistente en que los agentes se conviertan en kantianos⁴, y el último, según el cual los agentes se harían altruistas, parecen ser los menos problemáticos. Sobre este último cambio conviene no obstante hacer una matización. Una vez realizado el regateo, si todos somos altruistas es problema se resuelve bien. Lo mejor para ti es que yo colabore, de modo que haré A. Supuesto que el otro también es altruista, hará A también, de modo que la cooperación está garantizada. Pero que los agentes sean altruistas al realizar el regateo, en contra de lo que pueda parecer a primera vista, no cambia el carácter de este proceso. De todos los puntos óptimos posibles, yo prefiero los que te favorecen a ti y tu los que me favorecen a mí, de modo que, al igual que sucede entre jugadores auto-interesados, el problema tiene la forma de un juego de intereses opuestos, así como la situación previa al regateo no pierde su carácter de juego de intereses mixto.

Parfit matiza esta cuarta vía diciendo que los participantes no deben ser altruistas puros, e.d., personas que no tienen *ningún* interés por sí mismas, ya que de este modo se plantearía un dilema parecido.

Este dilema surgiría entre dos altruistas puros cuando se encuentran en situación de elegir entre a) darse a sí mismo un gran beneficio o b) dar al otro un beneficio más pequeño. Un altruista puro elegiría b, con lo cual el resultado sería peor para todos. Por ejemplo, consideremos la siguiente matriz

	B ₁	B ₂
A ₁	(1,1)	(0,9)
A ₂	(9,0)	(8,8)

Cada uno de los jugadores puede a) darle 1 al otro o b) darse 8 a sí mismo. Llamaremos A a la estrategia que da 1 al otro (e.d., las estrategias

⁴ Tal y como señala Parfit, el test kantiano, tomado literalmente, tiene sus propios problemas. El ejemplo habitual de estos es el de la práctica de una profesión, digamos la medicina. Yo no puedo racionalmente desear que todo el mundo practique la medicina, de modo que yo tampoco debo hacerlo. Y si todos somos kantianos, es de temer que las consecuencias fueran desastrosas. No obstante, el test puede refinarse lo suficiente como para impedir que surjan estos problemas. Por ejemplo, puede plantearse que la aplicación del test se limita a los casos en los que todos deseamos hacer algo, es decir, a los casos en los que la pregunta ¿que pasaría si todos hiciéramos lo mismo? surge de manera natural.

A_1 y B_1) y B a la que da β a uno mismo (A_2 y B_2). Si ambos eligen A , el resultado será $(1,1)$. Pero hay otro resultado que los dos preferirían a este, a saber (β,β) . Supongamos que los dos se ponen de acuerdo en no ayudarse y elegir ambos la estrategia β , con el fin de obtener ese resultado. Pero, de nuevo, este acuerdo no es estable. El jugador uno razonaría así. Si el jugador 2 mantiene el acuerdo y hace β , entonces es mejor para él que yo haga A . Y si no lo mantiene y hace A , entonces también es mejor para él que yo haga A . De modo que haré A . El jugador 2 haría un razonamiento similar y llegaría a la misma conclusión. Por tanto, sería $(1,1)$ el resultado.

Parece por tanto que la transformación de los agentes en altruistas puros no solucionaría los dilemas. Más bien, la cuarta solución moral deberá consistir en que los agentes sean lo suficientemente altruistas, no sobrepasando en ningún caso el límite de la benevolencia imparcial, e.d., una consideración imparcial de los intereses de todos, incluidos los propios.

Debe notarse que cualquiera de los cambios morales apuntados pueden dar lugar a cualquiera de las soluciones morales. Sería demasiado largo dar un ejemplo de cada motivación en relación con cada una de las soluciones, pero sí sería interesante ver algunos de ellos que pueden resultar más problemáticos.

Consideremos el primer cambio moral, consistente en que el agente se convierta en una persona de confianza. Debido al carácter de este cambio, puede pensarse que una persona así dispuesta elegirá siempre A con independencia de que A sea mejor o peor para él. Es decir, puede pensarse que un agente con esta disposición elegirá siempre A movido, no por la consideración de los beneficios que A le suponga, sino, simplemente, porque se ha comprometido a realizar A . Esto es en parte verdad. Un agente así dispuesto respetará el acuerdo *porque* se ha comprometido a hacerlo. Pero este "porque" no tiene necesariamente una lectura causal. Su lectura fundamental es condicional, e.d., un agente así dispuesto cumplirá el trato *si* se ha comprometido a hacerlo, puesto que es una persona de confianza. Pero puede suceder que un agente con esta disposición elija A *porque* para él ser un hombre de palabra tenga tanta importancia que, de no cumplir lo acordado, los beneficios conseguidos no compensen el haber

perdido su fama de hombre honrado, ante los demás o ante sí mismo². No voy a discutir si las personas de confianza dan lugar por lo general a las soluciones 4 y 5 y sólo excepcionalmente a la 3 o viceversa. La cuestión importante es que este cambio moral puede dar lugar a todas ellas.

Respecto a los cambios morales mediante los cuales los agentes pasan a ser reacios a ser aprovechados o se convierten en kantianos puede pensarse también que dan lugar a las soluciones 4 y 5. Sin embargo esto no es así, por el mismo motivo que no lo era en el caso anterior.

Acercas de el último cambio moral, por el contrario, puede pensarse que sólo puede dar lugar a la solución 3. Si alguien adquiere un interés suficiente por los asuntos de los demás, y más aun si llega al punto de conceder la misma importancia a estos que a los suyos propios, seguramente la estrategia A será mejor para él y, por esto, estará dispuesto a realizar A. Pero tampoco en esta ocasión sucede esto de manera necesaria. Desde luego, es indudable que nuestros intereses y preferencias, y, por tanto, nuestra función de utilidad, se alteran cuando empezamos a tomar en cuenta los intereses de los demás. Pero también es cierto que siempre hay un sentido en que nuestros propios intereses y los intereses de los demás son distintos, aunque nosotros estemos dispuestos a tomar en cuenta estos últimos.

Puede suceder que algunas personas tengan lo que podríamos llamar un "interés personal" por el bienestar de todo el mundo, es decir, que sientan hacia el bienestar de todos los demás el mismo tipo de interés que otro puede sentir por el de sus hijos, pareja, padres, amigos íntimos o por el de uno mismo. Para tales personas, su función de utilidad está determinada por estos intereses, del mismo modo que la mía lo está por mi interés hacia ciertas personas. También en estos casos es posible distinguir entre sus intereses y los de los demás. En parte esta distinción se hace debido a que, de hecho, la mayoría de estas personas tienen una mar-

² Tomada de modo literal, la presentación de las soluciones morales que hace Parfit parece obligar a esta posibilidad, ya que la solución 3 surge cuando el agente tiene una disposición a hacer A y esta disposición ha sido adquirida porque A es mejor para él. Sin embargo, creo que esta interpretación es errónea. No se trata de porque un agente haya adquirido una determinada disposición, sino de porque este dispuesto a hacer A en un momento determinado. Por tanto, la solución 3 se produce cuando un agente tiene una disposición a realizar A debido a que A, a raíz de un cambio interno del agente, es mejor para él.

cada tendencia a opinar sobre cuales son los intereses de los demás, separándose en este punto de la opinión de los propios interesados. Aparte de este hecho, la distinción se establece acudiendo a lo que serían los intereses de estas personas si estas no se preocuparan tanto de los demás, cosa siempre posible. Esta distinción es la que hace posible decir que alguien esta obrando en contra de su propio interés.

Por lo general, en estos casos estamos más dispuestos a hacer esta distinción de lo que lo estamos, por ejemplo, en el caso de una madre con su hijo. Cuando una madre se sacrifica por su hijo, incluso cuando llega a poner en peligro su propia vida, en cierto modo podemos decir que actúa contra sus propios intereses, pero realmente no consideramos que los intereses de su hijo no sean los suyos también. El motivo es que consideramos la preocupación de la madre como algo natural, o, si se prefiere, como algo común que nos ocurre a la gran mayoría de los seres humanos⁴. Aunque en menor grado, lo mismo ocurre con el interés de un hombre por sus amigos. No obstante, aquí no vamos a ocuparnos de si un interés personal por el bienestar de todo el mundo es natural o no. Lo cierto es que algunas personas (pocas, por cierto) lo sienten. Y cuando esto sucede, y estas personas escogen la estrategia A, se produce la solución 3.

Sin embargo, esto no es lo que sucede en todos los casos en los que el agente es lo suficientemente altruista para elegir A. Yo puede estar interesada en el bienestar de los demás, por ejemplo en el bienestar de los nicaragüenses, sin que esto se confunda en ningún momento con el interés personal que siento por un amigo. La diferencia es tremendamente difícil de señalar, pero existe.

Si yo tuviera que elegir entre salvar la vida de uno de aquellos o hacerle un regalo a mi amigo, elegiría lo primero. Pero si tuviera que elegir entre hacer un regalo a uno de los dos de tal modo que mi regalo les beneficiara a ambos igual, o incluso si beneficiara a mi amigo un poco menos, le haría el regalo a mi amigo. Si viera que alguien esta maltratando a mi amigo, saldría corriendo sin pensarlo dos veces, a cualquier hora y a cualquier lugar, para ayudarlo. Cuando veo un reportaje en

⁴ También en este caso hay matices. Habitualmente, no nos causa mucho problema afirmar que los intereses del hijo son los de la madre cuando el primero tiene 10 años. Cuando tiene 40 nos sentimos más inclinados a distinguir entre los intereses de ambos.

el que alguien esta siendo maltratado, me indigno e incluso, en ocasiones, pienso si debería correr a prestar mi ayuda, pero no lo hago. El reportaje acaba y empieza la película. Si la cosa era muy tremenda, me siento mal durante un rato, y es posible que me apunte a algún organismo internacional de ayuda. Esto indica una diferencia de grado.

Por otra parte, si tuviera que entregar algo valioso para sacar a un amigo de un llo, lo haría gustosa, mientras que si tuviera que hacerlo por un desconocido, aunque lo haría (al menos, lo haría si yo fuera una persona suficientemente altruista) no lo haría tan gustosamente. Me costaría más. Ante el agradecimiento de mi amigo, diría "de nada" de todo corazón. Ante el del extraño, estas palabras no pasarían de ser una frase hecha. Esto indica algo más que una diferencia de grado. Indica hasta que punto mi interés por el bienestar de mi amigo es personal, mientras que mi interés por el extraño no lo es.

Mi interés personal por mi amigo es algo "natural", un interés que yo, simplemente, tengo, como lo tengo, en mayor o menor grado, por mí misma. Mi interés no-personal por los extraños no es natural en este sentido. O, mejor dicho, si lo es un *cierto* interés. A mí no me da igual lo que les suceda a los demás. Prefiero que les vaya bien. Esto es un interés que yo tengo, igual que tengo interés en que determinado partido gane las elecciones o en que mi coche sea de un determinado color. Pero un interés que pueda ser calificado de "suficientemente altruista", e.d., que me haga estar dispuesto a elegir la estrategia A en vez de B, ya es otro asunto. Este tipo de interés sólo lo tengo si *me obligo* a considerar la situación desde un punto de vista imparcial. Este punto de vista siempre puede ser adoptado. Pero hacerlo o no es algo que depende de mi voluntad. No es el punto de vista que *yo tengo*. Es el punto de vista que *yo puedo adoptar* si me determino a hacerlo. Y puede ser que yo piense que *debo* adoptar ese punto de vista. Pero esto no tiene nada que ver con mis intereses personales. Es algo que *debo* hacer. Si lo hago, y desde ese punto de vista realizo mi elección, elegiré A. Pero lo elegiré tanto si A es mejor para mí como si no lo es. Puede ser que, al adoptar ese punto de vista, A ya no sea peor para mí. Pero también puede suceder que si lo sea. Y, a pesar de eso, elegiré A.

De modo que podemos concluir que el cambio moral consistente en que el agente pase a tener suficiente altruismo puede dar lugar tanto a la solución 3 como a la 4 y la 5.

APENDICE 10

De acuerdo con nuestra teoría, actuar de un modo moral no está racionalmente exigido. La manera más clara de expresar esto es la utilizada por Harsanyi, quien afirma que "si consideramos las reglas morales como posibles razones (e.d., motivos racionales) para que una persona realice las acciones ordenadas por esas reglas, las reglas morales sólo pueden ser interpretadas como imperativos hipotéticos"¹.

En tanto que las reglas morales son posibles razones para hacer algo, e.d., son posibles respuestas a la pregunta acerca de lo que debe hacerse en una situación determinada, son imperativos. Esto no hace sino recoger una característica de las normas morales universalmente reconocida, la prescriptividad. Podríamos entonces aceptar la definición dada por Kant de los imperativos "La representación de un principio objetivo, en tanto que es constrictivo para una voluntad, llámase "mandato", y la fórmula del mandato llámase "imperativo"². Sin embargo, podría pensarse que las reglas morales forman una clase de imperativos cuyas características lógicas son distintas a las del resto de los imperativos. Podría mantenerse, tal y como hace Kant, que mientras que el resto de los imperativos son hipotéticos, las normas morales se expresan en imperativos categóricos.

Por otro lado, la postura que aquí hemos adoptado y que se encuentra expresada en el texto de Harsanyi, mantiene que todo imperativo es hipotético. Es decir, los imperativos morales no serían distintos del resto de los imperativos. Es más, debemos mantener que los únicos imperativos posibles son los hipotéticos. Algo constituye una razón para realizar una acción (es decir, algo es un imperativo) si apela a una actitud, un deseo o un fin existente, y enuncia el comportamiento adecuado para alcanzar tal fin o ser coherente con tal actitud.

No obstante, al proponer la distinción entre imperativos hipotéticos y categóricos, Kant parece argumentar de un modo sólido. Creo por tanto conveniente dedicar este apéndice al análisis de ambas posturas con

¹ Harsanyi (1961), p.26.

² Kant (1811), p.40.

el fin de clarificar las razones existentes a favor de la postura que se defiende aquí. Por tanto, procederemos del modo siguiente. En primer lugar, analizaremos con cierta precisión en qué consiste la distinción kantiana y los argumentos que la apoyan y, en segundo lugar, analizaremos los argumentos a favor de nuestra propia postura.

Como es bien sabido, la diferencia fundamental entre ambos tipos de imperativos es que, mientras que los hipotéticos sólo prescriben una acción como buena en tanto que medio para conseguir alguna otra cosa, los categóricos la prescriben como buena en sí. Ahora bien, según la teoría kantiana los imperativos morales no pueden pertenecer al primer tipo por varios motivos:

1. porque lo importante desde el punto de vista moral no es la eficacia de un medio para conseguir un fin determinado, sino más bien que el fin en cuestión sea bueno y racional,

2. porque los imperativos morales mandan de manera incondicional y son universalmente válidos. Pero parece que no podrían tener estas características si fueran hipotéticos, y esto por dos motivos. En primer lugar, porque si los imperativos hipotéticos prescriben los medios adecuados para un fin, entonces su aceptabilidad está condicionada a la aceptación del fin por parte del agente. En este sentido, no ordenan una acción incondicionadamente, es decir, dependen de que exista una determinada inclinación en el agente mientras que los imperativos morales han de ser prescriptivos incluso contra la inclinación, y, por consiguiente, dejan abierta la posibilidad de actuar de un modo distinto a como ellos ordenan (no la posibilidad de hecho, puesto que esta existe en cualquier caso, sino, digamos, la posibilidad de derecho).

En segundo lugar, y por la misma razón, no ordenan universalmente. Es decir, su necesidad sólo existe en tanto que el agente acepta el fin. Ahora bien, como Kant argumenta, cualquier cosa alcanzable por medio de la acción humana puede ser propuesta como fin por cualquier sujeto, de donde se sigue por un lado que el número de este tipo de imperativos es infinito y, por otro, que no son prescripciones para todo agente sino sólo para aquellos que acepten el fin.

3. El argumento del punto 2 puede ser contestado de la siguiente manera. Kant admite la existencia de un fin compartido por todos los hombres, y no de forma accidental sino necesariamente: la felicidad. La coincidencia universal a la hora de aceptar la felicidad como fin es necesaria porque, en palabras de Kant, pertenece a la propia naturaleza del hombre que así sea. La existencia de este fin universal hace que la clase de los imperativos hipotéticos tenga que sufrir una división. Por un lado aparece la subclase de los imperativos problemáticos, que serían aquellos

que prescriben los medios necesarios para alcanzar un fin posible, es decir, un fin que el agente puede o no aceptar como meta de su acción, y por otro lado estará la subclase de los imperativos asertóricos, que serían los que prescriben las acciones necesarias para alcanzar un fin real, es decir, un fin efectivamente existente para todos. Este segundo tipo de imperativos estaría integrado por todos los que prescriben los medios para el fin de la felicidad, ya que este no sólo puede ser aceptado como tal, sino que lo es realmente y puede ser atribuido *a priori* a todos los hombres pues es una "necesidad natural" la que hace que sea un fin del hombre.

Una vez aceptada la existencia de la felicidad como un fin al que necesariamente tiende todo hombre, parece que las objeciones puestas al imperativo hipotético respecto a su carencia de incondicionalidad y universalidad desaparecen. Sin embargo, para Kant esto no sucede, puesto que

a) El concepto de "felicidad" es indeterminado. Esta indeterminación se da a dos niveles: primero, la indeterminación es debida a que cada hombre puede encontrar que tal o cual cosa concreta entra o no a formar parte de su felicidad y, segundo, incluso un mismo hombre puede no saber (y, según parece insinuar Kant, de hecho no sabe) *a priori* qué cosas van a proporcionarle la felicidad. Por estos motivos, el imperativo asertórico no prescribe incondicionada ni universalmente. No son incondicionados porque, pese a la necesidad del fin, no pueden exponer los medios necesarios para conseguirlo. Por ello no son siquiera propiamente mandatos, sino que tienen más bien el carácter de consejos que nos dicen lo que, según la experiencia, suele ser mejor medio por lo general para alcanzar la felicidad. Y no son universales porque dependen para su aceptación de una condición subjetiva contingente, a saber, que tal hombre particular encuentre tal y tal cosa entre aquellas que forman su felicidad.

b) Aún en el caso de que el concepto de "felicidad" no fuera indeterminado seguiría estando presente el motivo presentado en el punto 1. y que impide que los imperativos hipotéticos puedan ser imperativos morales. Es decir, aún en el caso de que pudiéramos decir con certeza absoluta "si quieres ser feliz, hazte funcionario", e.d., aunque este imperativo fuera una oración analítica y pudiéramos hacerla prescriptiva para todo ser racional, seguiría siendo cierto que se trata de un imperativo

que prescribe una acción en tanto que buena como medio y no en tanto que buena en sí misma.

Hasta aquí la teoría kantiana. Como hemos visto, uno de los principales motivos por los que se considera que los imperativos hipotéticos no pueden ser imperativos morales es que estos sólo prescriben una acción en tanto que es buena como medio. Sin embargo, este argumento se basa en parte en una mala interpretación de los imperativos hipotéticos. En efecto, tal y como dice Harsanyi

"Kant sólo considera imperativos hipotéticos donde la prótasis hace referencia a un fin hipotético mientras que la apódosis propone medios que conducen causalmente a ese fin, tal y como ocurre en el caso de las instrucciones técnicas, y tal y como también ocurriría en el caso de las reglas prudenciales si estas se interpretan tal y como lo hace Kant como instrucciones para producir la felicidad. Sin embargo, existen de hecho imperativos hipotéticos de uso común en los que la relación entre prótasis y apódosis es evidentemente *formal* y no *causal*; la prótasis especifica cierto criterio³ deseado mientras que la apódosis sugiere un objeto o un patrón de conducta que satisface ese criterio"⁴

Esta distinción no es nueva para nosotros. Como se recordara, en el capítulo VIII ya la habíamos utilizado al hablar de la relación entre querer el fin y querer los medios. Entonces utilizamos el ejemplo del paso a dos para aclarar que se quería decir con "relación formal", y decíamos que si alguien quiere bailar un paso a dos entonces necesariamente quiere bailar con otra persona, pero no porque al querer el fin haya que querer los medios, sino porque querer bailar un paso a dos es querer bailar con otra persona. La relación entre ambas cosas no es causal en absoluto. No es que el bailar con otra persona sea la causa de bailar en paso a dos, al igual que el que una figura geométrica tenga tres ángulos no es la causa de que sea un triángulo. Más bien, bailar con otra persona es bailar un paso a dos, al igual que tener tres ángulos es ser un triángulo. De modo que la oración "si quieres bailar un paso a dos, entonces tienes que bai-

³ "Criterio" no es, sin lugar a dudas, un término adecuado. Lo que aparece en la prótasis es, en todos los casos, un fin propuesto. El problema es que los términos "medio" y "fin" están fuertemente asociados con la relación causal, lo cual es sin duda el motivo de que Harsanyi intente evitarlos. Puesto que creo que la distinción es clara de todos modos, yo hablaré de fines y medios en todos los casos, señalando cuando sea necesario que se refiero a fines y medios relacionados de forma no causal. Sin embargo, siempre que se trate de una traducción, mantendré el término "criterio" del original.

⁴ Harsanyi (1961), p.24.

lar con otra persona" es un imperativo hipotético al igual que lo es la oración "si quieres bailar un paso a dos, entonces tienes que ir a una academia de baile", sólo que mientras que este es un imperativo de carácter causal, aquel es un imperativo de carácter formal.

Los imperativos³, sean causales o formales, pueden ser clasificados en dos grupos distintos según el sujeto cuyos fines están expresados en la prótasis. Desde este punto de vista, un imperativo será un "consejo" si sugiere una acción que conduce al cumplimiento de los fines del oyente, y una "demanda" si los fines a los que se hace referencia son los del hablante. Esta división no está tan clara como pudiera parecer a primera vista. Para que el imperativo resulte efectivo, e.d., para que pueda decirse que proporciona una razón para realizar determinada acción, es necesario que de algún modo haga referencia a los fines o actitudes del oyente⁴. Por tanto, dentro de las demandas los fines del oyente deben ser incluidos de algún modo, aunque la referencia principal siga siendo a los del hablante. Tal y como afirma Harsanyi, existen dos medios de apelación a los fines del oyente en el caso de las demandas. O bien apelamos a su interés benevolente para la satisfacción de los fines del hablante, o bien nuestras demandas están apoyadas en determinadas sanciones que pueden imponerse sobre el oyente si este incumple la demanda. Una formulación completa de la demanda tomaría en el primer caso una forma del tipo "si quieres hacerme un favor, haz X" mientras que en el segundo caso sería algo así como "si quieres que yo haga A (o deje de hacer B), haz X". Es fácil ver que las demandas que hacen referencia a sanciones son siempre de tipo causal, mientras que las otras son en la mayoría de los casos formales. En efecto, en el primer caso la acción del oyente es la causa de que la sanción se aplique, mientras que en el segundo caso el cumplimiento del imperativo no es la causa de que se haga un favor, sino que más bien es el favor mismo.

³ En adelante y salvo que se indique expresamente lo contrario, hablaremos simplemente de 'imperativos' sin hacer referencia a si son hipotéticos o categóricos, ya que estamos suponiendo que esta distinción no existe realmente y que todo imperativo es hipotético.

⁴ Qué esto es así puede comprobarse incluso a través de la propia formulación de los imperativos. Si intentamos expresar un imperativo que no apele a los fines del oyente sino sólo a los del hablante nos encontramos con algo así como "si quiero ir al cine, cómprame una entrada" lo cual se parece más a un sinsentido que a cualquier otra cosa.

Una vez que contamos con una teoría más adecuada acerca de los imperativos, podemos ver los motivos por los que las reglas morales pueden ser, y de hecho son, imperativos hipotéticos. Para empezar, veamos donde se sitúan las reglas morales dentro de este nuevo esquema. Podemos empezar por preguntarnos si los imperativos morales pueden ser demandas. Para ello, tenemos que preguntarnos si las reglas morales apelan a los fines del oyente de alguno de los dos modos empleados por las demandas. Respecto a la referencia a sanciones, parece indudable su ausencia en las reglas morales. Este tipo de reglas no consiguen nuestra aprobación mediante promesas de castigos y recompensas⁷.

No está tan claro a primera vista que no formen parte de las demandas que apelan a la benevolencia del oyente. Sin embargo la razón es sencilla. Simplemente, las reglas morales tienen carácter universal, mientras que no puede decirse lo mismo respecto a la benevolencia. De hecho, tal y como ya hemos dicho, es precisamente la inexistencia de la benevolencia universal lo que hace necesaria la moral. La regla moral que dice que no deben romperse las promesas es una regla aplicable a todos los casos en los que el agente haya prometido algo, y no sólo a aquellos en los que el agente sienta simpatía hacia el que ha recibido su promesa o preocupación personal por sus intereses. Y tampoco puede suponerse que la benevolencia universal sea un deber moral, puesto que ésta pertenece al campo de los sentimientos y no al de las acciones.

Por tanto las reglas morales han de ser algún tipo de consejo. Son consejos dirigidos a alguien que tiene cierta actitud moral, y consisten en decir qué tipo de conducta es coherente con tal actitud. Podemos preguntarnos ahora si las reglas morales son consejos de tipo causal o de tipo formal. Los imperativos morales tienen siempre un carácter formal. Mencionan en la prótasis cierta actitud moral que el oyente tiene y proponen en la apódosis el comportamiento que responde a esta actitud. La relación entre el comportamiento sugerido y la actitud no es de causa-efecto, sino más bien una relación entre un concepto y su definición parcial o total. Pensemos por ejemplo en el imperativo "si quieres ser gene-

⁷ Recordemos que precisamente la solución a los dilemas que surge a partir de la aceptación de reglas morales por parte de los agentes se distingue de las soluciones políticas precisamente porque estas son imponibles y aquellas no.

roso, haz X". Evidentemente, el que yo haga X no es la causa de que yo sea generoso. Más bien, hacer X es parte de lo que se entiende por "ser generoso" al igual que en el imperativo "si quieres jugar al fútbol, no toques la pelota con las manos a menos que seas el portero" el no tocar la pelota con las manos no es causa de que se juegue al fútbol, sino que es parte de la conducta adecuada para "jugar al fútbol".

Ahora que tenemos una teoría completa sobre los imperativos y que sabemos que tipo de imperativo son las reglas morales podemos revisar los argumentos de Kant.

1. El primer argumento de Kant está expresamente formulado contra los imperativos hipotéticos causales. Sin embargo, podría ser fácilmente reformulado de tal modo que cubriera también los imperativos no causales. Podría decirse que las reglas morales no pueden ser imperativos hipotéticos porque estos sólo prescriben una acción o bien en tanto que buena como medio para conseguir un determinado fin relacionado con aquel de modo causal o no causal. Por el contrario, las reglas morales se ocupan de la bondad de los fines.

No parece discutible que la moral se ocupa de la bondad de los fines y no de la eficacia de los medios. De todos modos, esto debe matizarse. Por supuesto que la moral se interesa por la eficacia de los medios², siempre que el fin sea moralmente bueno, y exige la búsqueda de los medios más adecuados. Querer el fin es estar dispuesto a poner los medios. Por ejemplo, no diríamos de nadie que intenta ayudar a alguien si no pone los medios para prestar tal ayuda. Esto sucede porque, en primer lugar, es imposible querer ayudar a alguien sin estar dispuesto a buscar los medios necesarios para hacerlo. En segundo lugar, porque resulta increíble que alguien este dispuesto a poner tales medios sin ponerlos, a no ser que tenga una explicación satisfactoria, del mismo modo que si vemos a alguien sentido que nos dice que quiere estar de pie no le creemos, a menos que este atado, que nos diga que el médico le ha prohibido levantarse o que sepamos que su voluntad es patológicamente débil. Y, en tercer lugar, y como consecuencia de lo anterior, porque lo que la moral

² Por lo general y salvo que queramos establecer una distinción entre imperativos causales y no causales, hablaremos de 'medios' en sentido general, abarcando con este término tanto los medios en sentido causal como los medios en el sentido de patrones de conducta que satisfacen un determinado modelo.

nos ordena es ser generosos u honrados y no "querer" serlo, ya que este "querer" no sería en realidad más que decir que se quiere.

Es decir, el que la moral se interese por la bondad de los fines no quiere decir que no se interese por la eficacia de los medios, aunque si quiere decir que no se interesa por la eficacia en sí.

Ahora bien, lo único que esto quiere decir es que no todos los imperativos hipotéticos son imperativos morales, verdad que no hay que preocuparse por defender ya que nadie ha mantenido lo contrario. En principio, no parece haber nada en contra de aceptar que las reglas morales puedan ser imperativos hipotéticos. Según la teoría propuesta por Harsanyi, serían reglas morales aquellos imperativos que expresan en la protasis una actitud moral.

2. El segundo argumento de Kant decía que las reglas morales, debido a su carácter incondicional y universal, no podían ser imperativos hipotéticos.

Atendamos primero a la incondicionalidad. Generalmente se admite que las reglas morales mandan incondicionalmente, pero en ocasiones no se deja claro qué es lo que se quiere decir con esto. Intentemos precisarlo mediante un ejemplo. Pensemos en la norma moral que nos dice que no debemos matar. Si decimos que esta norma es incondicional queremos decir que no esta condicionada a nada en su validez y que, por lo tanto, no admite excepciones. Sin embargo, dicho así parece que esta norma no es incondicional, ya que casi todos los códigos morales, incluida la moral del sentido común, reconocen la existencia de circunstancias que nos eximen de su cumplimiento, por ejemplo, cuando matamos en defensa propia como único medio para salvar nuestra propia vida.

Parte del problema es que hemos empleado el término "matar" y no otro como, por ejemplo, "asesinar". La diferencia entre estos dos términos es clara. "Matar" significa "producir la muerte de un ser vivo". Es decir, matar es algo que puede hacerse de muchos modos, en distintas circunstancias y contra seres muy distintos. Podemos ver que este término tiene dos características importantes: es un término con un nivel alto de generalidad y puede definirse en términos moralmente neutros, tal y como por ejemplo hemos hecho en la definición anterior. Por este motivo, si alguien nos dice que tal persona ha matado estamos interesados, antes de emitir un juicio moral, en saber a quién o en que circunstancias.

Sin embargo, podemos emplear palabras más precisas. Dependiendo de las circunstancias del caso en cuestión, podemos hablar de "homicidio" voluntario o involuntario, de "asesinato", de "sacrificio" etc. Cada uno de estos términos nos da una información sobre la acción que no nos da el término más general de matar. Fijémonos en "asesinar". Este término significa "matar a una persona sin que haya para ello ninguna justificación (moral)" (o, con más color, tal y como aparece en algunos diccionarios "matar traidora y alevosamente, a mansalva"). Por tanto, el término nos informa de cosas fundamentales para poder calificar moralmente la acción, tales como a qué tipo de ser se le ha dado muerte, o qué la muerte no se ha producido por accidente ni en defensa propia. En contraposición al término "matar", "asesinar" resulta mucho más preciso. Pero lo más importante es que esta precisión está realizada desde el punto de vista moral. (Desde luego, puede haber otros términos que realicen también algún tipo de precisión desde otro punto de vista, como por ejemplo "sacrificar".)

Podemos expresar la diferencia entre estos términos de un modo claro mediante la terminología utilizada por Kovesi en *Moral Notions*: "asesinar" es una noción completa, e.d., está enteramente formada desde el punto de vista moral, mientras que "matar" no lo es. Por esta razón, cuando sabemos que alguien ha matado no sabemos todavía lo suficiente para emitir un juicio moral sobre su acción, pero si se nos dice que alguien ha cometido un asesinato sí tenemos ese conocimiento. Naturalmente que aún no podemos precisar su grado de culpabilidad ni, por tanto, hasta qué punto se merece nuestra condena, ya que un asesinato puede cometerse en circunstancias atenuantes, como por ejemplo cuando el autor del crimen se encuentra bajo el efecto de alguna droga o sometido a una grave tensión emocional. Pero sabemos que, en el grado que sea, el autor es moralmente condenable, ya que el término no admite circunstancias eximentes, pues lo que indica precisamente es que tales posibles circunstancias no se han dado en este caso.

Muchos de los términos empleados en las reglas morales son nociones moralmente completas, pero hay algunos que no lo son. Un ejemplo de este segundo caso es "mentir". En un nivel muy general, distinguimos "mentir" de "no decir la verdad" y hacemos la precisión de que "mentir" significa "decir algo que no es verdad con intención de engañar". De este

modo, no decimos que los actores "mienten" en el escenario, ni que los niños "mienten" cuando están jugando a los médicos. Pero "mentir" sigue siendo un término tal que no nos permite pronunciarnos moralmente. Por ejemplo, la gente miente jugando al póquer (al menos los buenos jugadores) o se miente a un niño o a un enfermo grave o a un anciano para ahorrarle una preocupación innecesaria, etc.

Por ello, hay casos en los que una regla moral necesita precisiones suplementarias para poder decirse de ella que es incondicional, entendiendo, como hemos dicho, que "incondicional" quiere decir "sin excepciones". En los principios morales se nos dice que determinado tipo de acción es correcta o incorrecta, o que es obligatoria su realización o su no realización*. Ahora bien, parece que "correcto" e "incorrecto" funcionan de modo distinto cuando se aplican a términos moralmente completos y cuando se encuentran calificando a términos incompletos. En el segundo caso, su función es seleccionar la calificación moral de la instancia concreta del tipo de acción al que se refiere el término incompleto, mientras que en el primer caso actúan como "recordatorios"¹⁰. Es decir, no añaden ninguna precisión, sino que simplemente nos repiten lo que ya sabemos por el propio término, a saber, la calificación moral de la acción.

Una vez que tenemos principios morales que contienen solamente términos moralmente completos, o bien principios que incluyen todas las precisiones pertinentes, podemos decir que tenemos reglas incondicionales. Aquí "incondicionales" debe entenderse en otro sentido más concreto que el anterior "sin excepciones". En este nuevo sentido las reglas morales son incondicionales a dos niveles

a) A un nivel que podríamos llamar propiamente kantiano, son incondicionales en tanto que mandan sin tomar en cuenta las inclinaciones del sujeto, es decir, sin tomar en consideración si éste tiene algún deseo subjetivo hacia su cumplimiento o no.

b) A un nivel de mayor generalidad que incluye el anterior, son

* Todas las acciones moralmente incorrectas deben ser evitadas, pero no todas las acciones moralmente correctas constituyen un deber. Para simplificar, podemos suponer en este momento que todas las acciones correctas son deberes.

¹⁰ Ver Kovesi (67), IV, 2.

incondicionales en tanto que reclaman cumplimiento incluso cuando el curso de acción por ellas indicado entra en conflicto con otras reglas de carácter no moral. Estas reglas no morales pueden ser de distintos tipos y van desde lo que podríamos llamar reglas prudenciales a las leyes o las reglas de etiqueta. Es decir, las reglas morales tienen un carácter prevaleciente.

De hecho, si analizamos las reglas morales veremos que decir de ellas que son incondicionales no sólo es verdad, sino que resulta en cierto modo una afirmación de carácter analítico. En efecto, por un lado vemos que en las excepciones recogidas para precisar términos incompletos nunca se encuentran cláusulas del tipo "excepto cuando te apetezca" o "salvo cuando no hacerlo ofenda al buen gusto". Y en las nociones moralmente completas vemos que su significado incluye una referencia al efecto de señalar que las únicas excepciones admitidas deben tener una justificación moralmente independiente, como cuando decimos que robar es incorrecto pero admitimos que está permitido cuando es la única manera de salvar una vida humana. Es decir, en su sentido está incluido el que sólo pueden dejar de cumplirse cuando entran en conflicto con otra regla moral de carácter superior.

Por otro lado, podemos comprobar la analiticidad de la afirmación acerca de la incondicionalidad de las reglas morales si nos fijamos en que ser incondicional es uno de los rasgos que una regla debe presentar para que podamos reconocerla como regla moral. Si estamos haciendo una recopilación de las normas de conducta de una sociedad con el fin de seleccionar las reglas morales mantenidas por la comunidad no admitiríamos en esta lista ninguna regla cuyo cumplimiento dependiera de las inclinaciones del agente o que pudiera ser relegado por otro tipo de reglas¹¹.

La cuestión por tanto es si los imperativos hipotéticos pueden o no mandar incondicionalmente. Un imperativo hipotético nos informa de lo que tenemos que hacer para conseguir un cierto fin, o para adecuar nues-

¹¹ Naturalmente, no quiero decir que esta sea el único medio de reconocimiento de las normas morales, ni siquiera el más importante. Lo único que quiero decir es que esta característica de la incondicionalidad debe estar presente.

tra conducta a un determinado modelo¹², siendo este último el caso de los imperativos morales. Son posibles muchos imperativos morales y, por tanto, muchos los modelos propuestos: unos nos informan de la conducta requerida para ser generoso, otros de lo que debe hacerse para ser valiente, honrado o buen ciudadano. A pesar de esta diversidad existe en todos los casos un denominador común. En todos ellos se nos indica el comportamiento moralmente correcto.

Podría decirse que entonces los imperativos morales no serían incondicionales, pues sólo mandan en el caso de que se quiera seguir una conducta moral. Desde luego, esto es una condición, pero de ningún modo es una condición en el sentido en que lo son aquellas que hemos descartado del ámbito de la moralidad. No nos dice en qué casos debemos seguir el comportamiento requerido, sino que más bien indica que esa conducta debe seguirse siempre que la ocasión se presente, nos informa de una conducta que debe ser seguida incondicionalmente: tanto si nos apetece como si no, sean cuales sean nuestros intereses personales y nuestros fines particulares, sean cuales sean nuestras relaciones con las personas afectadas por la acción. La "condición" presentada por los imperativos morales sólo se refiere a la propia condición de la moralidad, es decir, sólo establece la condición de que queramos decidir nuestro curso de acción desde el punto de vista moral. El punto de vista moral es incondicional. Si se quiere adoptar este punto de vista, se han de seguir sus reglas de forma incondicional. Pero esto no significa que los mandatos morales manden al que no quiera adoptar su punto de vista. Los imperativos morales son incondicionales porque para satisfacer el concepto explicitado en su prótasis hay que cumplir incondicionalmente la conducta indicada en la apódosis. Pero no porque haya que cumplirlos si no se desea satisfacer ese modelo.

En este sentido, podemos seguir diciendo que los imperativos morales presentados como imperativos hipotéticos son incondicionales, pero es necesario recordar que significa la incondicionalidad de las normas morales.

¹² De ahora en adelante, al referirme a los fines propuestos por los imperativos morales hablaré de 'modelos' o 'modelos de conducta' para distinguirlos de otros fines y señalar la relación no causal que se da en este tipo de imperativos.

Al llegar a este punto vemos que el problema se traslada a la cuestión de la universalidad. Al igual que en el caso de la incondicionalidad, se reconoce generalmente que los principios y juicios morales son universales. Veamos en qué sentido se admite esto.

Habitualmente se define la universalidad como la propiedad que tienen los juicios y principios morales de ser aplicables a toda situación relevantemente similar. Qué es o qué no es "relevantemente similar" viene determinado por la propia regla que se está aplicando. Tomemos un ejemplo simple, el juicio "asesinar es incorrecto" y la correspondiente regla moral "no se debe asesinar". Esta regla y este juicio son aplicables a todos los casos relevantemente similares. Ahora bien, desde el punto de vista de la regla, los aspectos relevantes de un caso son aquellos que permiten su aplicación, e.d., los casos de asesinato. "Asesinato" es un concepto moral y un concepto moral es, utilizando las palabras de Brennan¹³, una clasificación de acciones que se juzgan correctas o incorrectas por el mismo motivo y, por medio de tales conceptos, somos capaces de reconocer la similitud de dos casos, por lo demás enteramente distintos, como instancias de lo mismo.

En este sentido, la universalidad es una exigencia de consistencia lógica. En efecto, si juzgamos que X es incorrecto porque es un caso de asesinato, estamos obligados lógicamente a juzgar incorrectos todos los casos en los que esté presente la misma razón que nos hizo emitir el juicio anterior, e.d., todos los casos de asesinato. La universalidad es por tanto un principio formal que tiene que ver con la corrección o validez del razonamiento moral.

La universalidad así entendida tiene mucho que ver con la incondicionalidad, pues significa que la validez de un juicio o la aplicación de una regla sólo puede impugnarse por referencia a los factores relevantes de la situación, e.d., que cuando una regla es aplicable a un caso, entonces debe seguirse incondicionalmente.

Los imperativos morales en forma hipotética son universales en este sentido. La prótasis de estos imperativos no nos dice *cundo* es aplicable la regla. Las condiciones de aplicabilidad de ésta están expresadas en ella misma del modo antes indicado. Lo que indica la prótasis es

¹³ Ver Brennan (1977), 1, 7.

el modelo que estaremos cumpliendo si seguimos la regla. Pero seguir la regla *significa* aplicarla de forma universal. Esto puede verse con claridad en un ejemplo ofrecido por Hare

"Supongamos que yo le digo a alguien "Usted no debe fumar en este compartimiento", y que en el compartimiento hay niños. La persona a la que me dirijo, si se pregunta por qué yo he dicho que no debía fumar, probablemente mirará en derredor, observará a los niños y comprenderá la razón. Pero supóngase que, habiendo examinado todo lo examinable en el compartimiento, dice "Esta bien; me iré al de al lado; allí hay otro compartimiento tan bueno como éste; de hecho es exactamente igual, y hay también niños". Al oír esto yo pensaría que no ha comprendido la función de la palabra "debe", porque "debe" siempre se refiere a un principio general; y si el compartimiento contiguo es en realidad exactamente como éste, todo principio aplicable a éste lo debe ser al otro. Yo podría por tanto responder: "Pero mire, si usted no debe fumar es éste compartimiento y el otro es justamente como este, tiene la misma clase de ocupantes, las mismas advertencias en las ventanillas, etc., entonces obviamente no debería fumar tampoco en ese otro"¹⁴

Hare habla de dos compartimiento exactamente similares. Pero naturalmente no es necesario que se de ese grado de similitud. Tal como hemos dicho, basta con que ambos sean relevantemente similares. Obviamente, los aspectos relevantes en este caso son precisamente los mencionados: que los ocupantes sean del mismo tipo, que en ambos exista una prohibición o, quizá, que el sistema de ventilación sea similar o que los materiales en ambos casos sean fácilmente combustibles. Otras cosas, como el color de los asientos o la numeración del compartimiento son factores irrelevantes para la cuestión de si se debe o no fumar en ellos. Naturalmente, la persona que escucha la prohibición no tiene porque saber cuales son los factores relevantes en un caso concreto. Puede suceder que el oyente piense que el motivo por el que no se debe fumar en este compartimiento es porque tiene la tapicería verde. En nuestro ejemplo es difícil que esto suceda, y si alguien nos dijera esto probablemente pensaríamos que nos está tomando el pelo, pero puede suceder y, en otros ejemplos, con mucha mayor plausibilidad. Pero la cuestión no es esa. La cuestión es que si el oyente conoce la razón o razones por los que no se debe fumar en este compartimiento, entonces sabe que tampoco debe hacerlo en los casos en los que estas mismas razones están presentes.

¹⁴ Hare (1978), pp.176-7.

Sin embargo, Kant habla de universalidad también en otro sentido. En este segundo sentido, una regla es universal cuando se dirige a todo ser racional en tanto que tal, sin tomar en consideración sus peculiaridades.

De nuevo resulta necesaria una distinción similar a la que acabamos de hacer al hablar de cuál es el sentido de la condición expresada en la prótasis de los imperativos morales. Esta distinción estaría entre a) cuál debe ser la conducta de un ser racional ante la regla moral y b) la cuestión de si la condición de ser racional fuerza a la utilización de la regla como principio de conducta.

a) Hemos dicho que la aplicación universal de la regla es un requisito de coherencia lógica. Por tanto, en este sentido al aceptar una regla estamos forzados, en tanto que racionales, a aplicarla de modo universal. También hemos dicho que la propia regla indica en qué circunstancias es aplicable. De modo similar, también indica para quién es aplicable. Esta indicación viene dada por la ausencia de palabras que puedan interpretarse como nombres propios o descripciones definidas en la formulación de la regla. Es decir, en la especificación de los factores relevantes que hacen aplicable la regla no se encuentran expresiones de ese tipo. Una regla moral debe ser universal en este sentido, y esto viene exigido también por la racionalidad de las reglas. De igual modo que el juicio "es incorrecto asesinar" es aplicable a cualquier caso de asesinato, independientemente de la circunstancia de que este se cometa en enero o en marzo, también es aplicable con independencia de que lo cometa no u otro sujeto individual, pues ambas cosas, el mes o el autor, resultan irrelevantes desde el punto de vista de la regla. Podemos decir, siguiendo la distinción de Kovesi¹⁸, que las reglas morales, más que dirigirse a todos, se dirigen a cualquiera. Su mandato de, por ejemplo, no cometer asesinato, está dirigido a cualquiera. La universalidad de las reglas morales no consiste en que todos deben realizar determinada acción, sino en que esta acción se debe realizar siempre en determinadas circunstan-

¹⁸ Ver Kovesi (67), II, 3.

cias características¹⁴.

Resumiendo este punto, seguir una regla moral implica seguirla universalmente, aplicándola a todos los casos relevantemente similares. Los imperativos morales en formulación hipotética conservan este tipo de universalidad.

b) Este sentido de universalidad es el que resulta más cuestionable. También es a este sentido al que parece referirse Kant prioritariamente. "Universalidad" se refiere ahora no sólo a que debemos aplicar la regla universalmente sino a que debemos aplicarla en tanto que somos racionales. Es decir, no se refiere al modo de aplicación de la regla, sino más bien a que esas reglas son tales que todos debemos aplicarlas.

Es importante notar que la cuestión no es si debemos o no aplicar las reglas morales, e.d., si debemos o no aceptarlas, sino a si esto es o no un requisito de racionalidad, tal y como Kant pretende. Es en este punto donde aparece la diferencia entre la postura kantiana y la que mantiene que los imperativos morales son hipotéticos. Los imperativos en forma hipotética se dirigen a aquellos que quieren adecuar su comportamiento al cumplimiento de determinados modelos, mientras que las formas categóricas propuestas por Kant se dirigen a todos. ("Dirigirse a " debe entenderse en este contexto como "resultar obligatorias para".)

Kant intenta probar que las leyes morales son universales en este sentido acudiendo para ello a la condición de ser racional, y afirma que es en tanto que somos racionales que estamos obligados a seguirlas. Por el contrario, los que entienden las reglas morales como imperativos hipotéticos mantienen que "es imposible probar que alguien actúa de forma irracional por no conformar su comportamiento a determinado criterio formal, si ese alguien simplemente no quiere conformar su conducta a ese criterio"¹⁷. Es decir, si alguien tiene una actitud cívica y quiere ac-

¹⁴ El por qué las reglas morales son universales en este sentido es otra cuestión y tiene que ver con las razones que estamos dispuestos a aceptar como justificantes de las reglas morales y con el activo que nos hace estar así dispuestos. Habitualmente, una de las cosas que exigimos es que no sean arbitrarias, de modo que una regla del tipo "No se debe asesinar salvo los martes" o "Nadie debe asesinar menos Federico" serían difícilmente aceptables, a no ser que creamos, justificadamente o no, que ser martes o ser Federico son un exigente no arbitrario. Que lo creamos es improbable. Que nuestra creencia este justificada, imposible.

¹⁷ Harsanyi (76), p.30.

tuar de modo que su comportamiento satisfaga los requisitos de civismo, podremos mostrarles que ser cívico implica, entre otras cosas, pagar los impuestos y, por tanto, que el mandato de pagarlos es obligatorio para él si quiere satisfacer tal modelo, si como que al adoptar la regla "deben pagarse los impuestos" se compromete a aplicarla de modo universal y a juzgar como conducta incorrecta toda conducta que viole esta regla. Y podemos decir que todo esto es obligatorio para ese alguien en tanto que es un ser racional.

El problema es qué ocurre si alguien no desea ser cívico. Desde luego, de nada vale mostrarle que ser cívico implica comportarse de determinada forma. El podría estar perfectamente de acuerdo con nosotros sin pensar que el asunto le concierne. Su carácter racional se mostrará en ese acuerdo. Podría acusársele de irracionalidad si juzgara correcto que los demás paguen sus impuestos, pero que resulta correcto que él no los pague. Podemos hacerle ver que su actitud resulta incoherente con su vida, porque él se aprovecha de los impuestos que paguen los demás, admitiendo implícitamente que los demás hacen muy bien en pagarlos. Pero si nos dice que no cree que hagan bien, sino más bien que son tontos, pero que, ya que lo hacen, él se aprovecha, podemos acusarle de fresco, de inmoral y de algunas otras cosas, pero no de irracional. El se comporta de la manera adecuada para satisfacer el modelo de incivismo que se propone y no juzga correcta en los demás la postura contraria ni sostiene la validez de una regla en forma no universal.

Esto no quiere decir que no podamos hacer nada para modificar su actitud. Podemos mostrarle las consecuencias de la violación general de la regla a fin de que él pueda reflexionar sobre si realmente quiere vivir en un mundo así, o podemos apelar a sus sentimientos de simpatía hacia los demás con el fin de mostrarle lo que supondría para mucha gente la instauración de un régimen de vida en el que esté ausente la solidaridad social. Pero todos estos argumentos están dirigidos a cambiar su actitud y no a mostrarle la irracionalidad de su conducta.

Naturalmente, todo esto depende de la teoría de la racionalidad individual de la que hemos partido. Según esta, la elección de la estrategia altruista sólo es racional si estamos personalmente interesados en ella, de tal modo que actuar según esa estrategia maximiza nuestra utilidad. Si no es así, es decir, en los casos en los que se presenta un au-

téntico dilema, la elección de la estrategia altruista no sólo no es racional sino que es irracional. Puede pensarse que este concepto de racionalidad no recoge el requisito de universalidad. Pensemos por ejemplo en el dilema del prisionero. Por las razones repetidamente expuestas, cada uno de los individuos actuara de forma racional si confiesa. Esto es lo mejor para él bajo cualquier supuesto acerca de la conducta del otro. Imaginemos una regla moral del tipo "no confesar". Si ellos adoptan esta regla, entonces la aplicarán universalmente. La cuestión es si racionalmente estamos obligados a aceptarla. Kant podría argumentar diciendo "Bien. Seguro que cada uno de ellos quiere que el otro no confiese. Es decir, cada uno desearía que la regla se aplicara. Pero entonces tiene que aplicarla el mismo, porque no se puede desear la aplicación (necesariamente universal) de una regla sin aplicarla uno mismo."

Este argumento no es válido. Lo que cada uno desea es que el otro confiese. Si uno pudiera hipnotizar al otro a distancia y obligarle a confesar, lo haría encantado. Entonces él no confesaría y el resultado sería para él el mejor posible. Pero esto no es desear que el otro actúe racionalmente, sino que es precisamente desear que el otro actúe irracionalmente. Desde luego, si uno tuviera que elegir entre un mundo en el que todos respetaran la regla y un mundo en el que no se respetara, elegiría lo primero. Esta elección es racional. Pero los prisioneros en sus celdas no tienen que elegir un mundo, sino una estrategia a seguir en una situación dada. Si el que yo siguiera la regla garantizará su aplicación universal, lo racional sería seguirla. Pero no es así. Sin duda, elegir moralmente es elegir como si se estuviera eligiendo un mundo. Pero puesto que no estoy eligiendo un mundo (porque he de actuar como si lo estuviera haciendo) Si lo hago, mi conducta es irracional.

Si el test de universalizabilidad kantiano es parte de lo exigido para comportarse moralmente (como muy probablemente lo sea), entonces actuar según una norma de conducta que no pase este test es actuar de modo inmoral. Pero de ningún modo es actuar irracionalmente. Es irracional desde el punto de vista moral, pero desde el punto de vista de la elección individual no lo es en absoluto.

En otras palabras, parece que, si bien puede decirse que hay una manera racional de elegir moralmente, no parece que pueda mostrarse que sea racional elegir moralmente. En muchas situaciones, a saber, en las

situaciones de dilema, sabemos que todos estaríamos mejor si todos nos comportáramos moralmente. Y cada uno de nosotros estaría mejor todavía si los demás se comportaran moralmente. ¿Puede un agente racional desear que los demás se comporten moralmente? Por supuesto. No hay nada incoherente en este deseo. La pregunta de Kant es distinta. Kant pregunta si puedo desear que nadie se comporte moralmente. ¿Puedo desear que todo el mundo mienta? La respuesta es claramente negativa. "Pero entonces" diría Kant "yo no puedo mentir". Esto no es verdad. Si yo deseo comportarme moralmente, entonces tengo que seguir unas reglas de carácter universal. Pero no hay nada contradictorio en el deseo de que los demás no mientan. Sería contradictorio querer que nadie mienta y querer al mismo tiempo mentir yo. Pero no lo es querer que no mientan los demás. Es contradictorio querer aplicar una regla universal y no aplicarla universalmente. Pero no lo es querer que los demás la apliquen.

Harsanyi compara las reglas morales con las reglas de otras teorías normativas, tales como la teoría de la decisión y la teoría de juegos. Todas ellas se expresan en imperativos hipotéticos de tipo no causal. En todos estos casos puede hacerse una distinción entre reglas de orden inferior y reglas de orden superior. Fijémonos por ejemplo en las reglas de la teoría de la decisión. Tenemos por un lado reglas del tipo "si quieres maximizar la utilidad esperada, haz x". Paralelamente, están las reglas morales del tipo "si quieres maximizar la utilidad media de tu sociedad, haz X". Estas reglas de orden inferior se siguen de dos premisas, una de las cuales es una regla superior y otra que es una oración empírica. Las reglas de carácter superior son verdades analíticas, en las cuales la apódosis expresa la conducta implicada por la prótesis. Las utilizadas para derivar estos imperativos son, "si quieres comportarte racionalmente, entonces maximiza la utilidad esperada" y "si quieres comportarte como la haría un espectador simpático e imparcial, maximiza la utilidad media" respectivamente¹⁰.

¹⁰ Puesto que hemos defendido que la elección moral puede identificarse con la elección de un espectador simpático e imparcial, podemos suponer que hay aun una regla superior que diría "si quieres realizar tus elecciones desde el punto de vista moral, elige tal y como lo haría un espectador simpático e imparcial". Esta regla sería una oración analítica, al menos si consideramos que una elección moral es una elección (imparcial) entre los intereses de todos, como hemos mostrado en el capítulo I. Alguien puede pensar que no es así, sino que elegir moralmente es, por ejemplo, elegir de tal modo que se sea fiel a la patria. Ese alguien tendría que admitir que la elección moral no es una elección imparcial. De hecho,

La cuestión es que no sólo no hay ninguna exigencia racional de ser morales, sino que incluso en muchas ocasiones hay una exigencia racional de no serlo. Es decir, en muchas ocasiones el imperativo "si quieres comportarte racionalmente (o, si queremos, si eres racional), comportarte como lo haría un espectador simpático e imparcial" es un mal imperativo. En tanto que comportarse como un espectador imparcial no equivale a maximizar la utilidad esperada (que es la conducta que satisface el requisito "ser racional"), esta regla no puede ser considerada como regla de carácter superior. Puede que en ocasiones comportarse como un espectador imparcial maximice de hecho la utilidad esperada. En estos casos, esta es la conducta racional. Y cuando no es así, entonces simplemente es un imperativo tan malo como "si quieres dormir bien, tomate un litro de café fuerte antes de irte a la cama".

tal y como dice Harsanyi "todos los códigos de conducta no humanistas son meramente expresiones de preferencias personales contingentes - aunque posiblemente muy desinteresadas- por parte de la gente que adopta tales códigos, mientras que el código del humanismo imparcial simpático es el único que por definición da el mismo peso a las preferencias de las demás personas" (Harsanyi (76), p.35). Esto es motivo suficiente para considerar que este código es el único que merece el calificativo de moral, al menos si entendemos la moral tal y como lo estamos haciendo nosotros. En cualquier caso, en tanto que esta discusión es terminológica no tiene demasiada trascendencia.

BIBLIOGRAFIA

- ARISTOTELES (1981). *Ética a Nicomaco*. Centro de estudios Constitucionales, Madrid.
- AXELROD, R. (1986). *La evolución de la cooperación*, Alianza Universidad, Madrid.
- BAIER, K.. (1985), "Maximization and Fairness", *Ethics*, Vol. 96.
- BENTHAM, J..(1982). *An Introduction to the Principles of Morals and Legislation*, Methuen, New York.
- BOARDMAN, W.S. (1987). "Coordination and the moral obligation to obey the law", *Ethics*, Vol. 97
- BRANDT, R. (1979). *A Theory of the Good and the Right*, Clarendon Press, Oxford.
- BRAYBROCKE, D. (1987). "Social Contract Theory's Fanciest Flight", *Ethics*, Vol. 97
- BRENNAN, J.M. (1977). *The Open-Texture of Moral Concepts*, The MacMillan Press LTD, London-Basingstoke.
- DAYTON, E. (1979). "Utility Maximizers and Cooperative Undertakings", *Ethics*, Vol. 90.
- ELSTER, J. (1979). *Ulysses and the Sirens*, Cambridge U.P., Cambridge.
(1985). "Rationality, Morality and Collective Action", *Ethics*, Vol. 96
- FISHBURN, P.C. (1968). "Utility Theory", *Management Science*, Vol. 14
- GAUTHIER, D. (1986). *Morals by Agreement*, Clarendon Press, Oxford.
(1987). "Reason to be Moral?", *Synthese*, Vol. 72
- HAHN, F. & HOLLIS, M. (eds.) (1979). *Philosophy and Economic Theory*, Oxford U.P., Oxford
- HARE, R.M. (1978). *The Language of Morals*, Oxford U.P., Oxford.
(1981). *Moral Thinking*, Clarendon Press, Oxford.
- HARSANYI, J.C. (1956) "Approaches to the Bargaining Problem before and after the Theory of Games: a critical discussion of Zeuthen's, Hicks' and Nash's theories", *Econometrica*, n24.
(1975). "The Tracing Procedure", *International Journal of Game Theory*, 4.
(1976). *Essays on Ethics, Social Behavior, and Scientific Explanation*, D. Reidel Publishing Company, Dordrecht-Holland/Boston, U.S.A.
(1977a). "Morality and the Theory of Rational Behavior", *Social Research*, Vol.44, n24.

- (1977b), "On the Rationale of the Bayesian Approach", en Butts and Hintikka (eds) *Fundational Problems in the Special Sciences*, D. Reidel P.C., Dordrecht-Hollan, pp.381-392.
- (1977c), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge U.P., Cambridge.
- (1983), "Bayesian Decision Theory. Subjective and Objective Probabilities, and the acceptance of Empirical Hypotheses", *Synthese*, Vol. 57
- (1985) "Does Reason tell us what Moral Code to follow and, indeed, to follow any Moral Code at all?" *Ethics*, Vol.96.
- (1987), "The Tracing Procedure: a self-correcting reasoning procedure", *Theory and Decision*, Vol.23
- HOLLIS, M. (1987), *The Cunning of Reason*, Cambridge U.P., Cambridge.
- HUME, D. (1981), *Tratado de la naturaleza humana*, Editora Nacional, Madrid.
- KAGAN, S. (1986), "The Present-day Theory of Rationality", *Ethics*, Vol. 96.
- KANT, E. (1981), *Fundamentación de la metafísica de las costumbres*, Espasa-Calpe, S.A., Madrid.
- KOVESI, J. (1967), *Moral Notions*, Routledge & Kegan Paul LTD, London.
- KRAUS, J.S. & COLEMAN, J.L. (1987), "Morality and the Theory of Rational Choice", *Ethics*, Vol. 97
- LOCKE, J. (1963), *Two Treatises of Government*, en *The work of John Locke*, Vol. V, Scientia Verlag Aalen, Germany.
- LUCE, R.D. & RAIFFA, H. (1957), *Games and Decisions*, John Wiley & Sons, Inc. New York . London . Sidney.
- MCPHERSON, M.S., (1982), "Mill's Moral Theory and the Problem of Preference Change", *Ethics*, Vol. 92
- MENDOZA, J. (1986), "Parfit on Directly Collectively Self-Defeating Moral Theories", *Philosophical Studies*, Vol. 50
- (1987), "Gauthier's "Morals by Agreement" and the two Kinds of Rationality", *Ethics*, Vol. 97
- MILL, J.S. (1974), *El Utilitarismo*, Aguilar Argentina S.A., Buenos Aires.
- MORSIMORE, G. (ed.), (1971), *Weakness of Will*, MacMillan & Co. LTD, London-Basingstoke.
- NAGEL, T. (1970), *The Possibility of Altruism*, Princeton U.P., Princeton, New Jersey.
- NASH, J.F. (1950), "The Bargaining Problem", *Econometrica*, n218.
- NORMAN, R., (1971), *Reasons for Actions*, Basil Blackwell, Oxford.

- NOZICK, R. (1974). *Anarchy, State and Utopia*, Basil Blackwell, Oxford.
- LEACH, J. (1977). "The dual Function of Rationality", en Butts & Hintikka (ed.) *Foundational Problems in the Special Sciences*, D. Reidel Publishing Company, Dordrecht-Holland.
- PARFIT, D. (1979). "Prudence, Morality and the Prisoner's Dilemma", *Proceedings of the British Academy*, Vol. 65.
(1986). *Reasons and Persons*, Oxford U.P., Oxford.
(1986b). "Comments", *Ethics*, Vol.96.
- RAWLS, J. (1977). *A Theory of Justice*, The Belknap Press of Harvard U.P., Cambridge, Massachusetts.
- REGAN, D. (1980) *Utilitarianism and Co-operation*, Clarendon Press, Oxford.
- SEN, A. (1976). *Elección colectiva y bienestar social*, Alianza Universidad, Madrid.
(1974). "Choices, Orderings and Morality", KÖRNER, S. (Ed.), *Practical Reason*, Blackwell, Oxford.
- SEN & WILLIAMS (1982). *Utilitarianism and beyond*, Cambridge U.P., Cambridge.
- SIDGWICK, H. (1981). *The Methods of Ethics*, Hackelt Publishing Company, Cambridge.
- SIMON, H.A. (1955). "A Behavioral Model of Rational Choice". *Quarterly Journal of Economics*, Vol.69.
- SLOTE, M. (1985). *Common-sense, Morality and Consequentialism*, Routledge and Kegan Paul, London-Boston.
- SOWDEN, L. (1983). "That there is a Dilemma in the Prisoner's Dilemma", *Synthese*, Vol. 55.
- TULLOCK, G. (1967). "The Prisoner's Dilemma and Mutual Trust", *Ethics*, Vol. 77.
- WILLIAMS, B., (1981). *Moral Luck*, Cambridge U.P., Cambridge.

INDICE DE AUTORES

Aristóteles 194, 195, 284-286, 292

Axelrod 146

Bentham 136

Brandt 22

Brennan 359

Elster 54, 187-189, 191, 193, 200-202, 204

Gauthier 16, 17, 41, 43, 56, 87, 89, 90, 93, 94, 96, 98, 99, 109, 110, 113, 114, 116-119, 123, 124, 127, 131-133, 145-151, 153-155, 156-163, 165-168, 298, 304-306, 318-331, 333-335

Hare 32, 360

Harsanyi 3, 41-43, 49, 59, 80-82, 121, 122, 127, 170, 175, 176, 180, 181, 239, 274, 305, 306, 308, 313, 315-317, 319, 329-331, 346, 350, 351, 354, 362, 365, 366

Hollis 45

Hume 164

Kagan 221, 222, 224, 225, 234-236, 242

Kant 153, 346, 348-350, 353, 354, 361, 362, 364, 365

Kovesi 355, 356, 361

Locke 93, 333, 334

Luce 37, 42, 59, 64, 68, 78, 85, 121, 129

Mill 92, 136, 141

Nozick 334

Parfit iii, 23, 25, 26, 67, 69, 135-137, 144, 164, 165, 174, 175, 186, 187, 207, 208, 210, 211, 216-222, 225-228, 234, 236-246, 249, 250, 251-253, 255, 256, 258-262, 336, 337, 340, 342

Sen 14-16, 18, 19, 268-270, 272, 294, 295, 309

Williams 18

INDICE DE MATERIAS

- acuerdos iii, 70, 72, 73, 88, 90, 91, 93, 96-98, 101, 107-110, 113, 114, 116-124, 127, 129, 131-136, 145, 146, 268, 274, 279, 280, 322, 339
 - auto-imponibles 133
 - imponibles 123, 131, 133-135, 146, 268, 274, 352
 - insatisfactorios iii, 90, 98, 114, 119, 120, 123, 131
 - satisfactorios iii, 90, 91, 93, 98, 101, 107-110, 116, 118, 120, 122-124, 127, 131, 132, 133, 145, 146
- altruismo 157, 163-165, 172, 175, 260, 345
 - particular 10, 11, 172, 240, 283, 349
 - paternalista 173, 175
 - universal 172, 173, 175, 179
- aprovechado 93, 94, 97, 110, 111, 131, 337
- cambios de preferencias 190
 - endógenos 190, 191, 193, 194, 198, 199, 201
 - exógenos 190, 193, 199
 - temporales inconsistentes 190, 193, 199
- condiciones de negociación 93, 102-112, 114-119, 145, 146
- condiciones de racionalidad 13, 43-45, 57, 58, 63, 64, 192, 226, 309
 - completud 14, 16-19, 229, 294
 - continuidad 42, 304-306
 - ordenación 14-16, 19-21, 23, 28, 30-32, 39, 41, 45, 47-49, 51, 63, 64, 156, 177-179, 186, 189, 197, 204, 229, 230, 269, 270-273, 294-298, 304
 - reflexividad 14-16, 229, 294
 - sure-thing principle* 41
 - transitividad 14, 15, 19, 20, 229, 295, 296, 298
- dilemas i, 69, 139, 146, 147, 157, 162-164, 168, 169, 171-175, 179, 183, 216, 217, 256-259, 265, 269, 275, 278, 279, 341, 352
 - del prisionero 66, 68, 72, 135, 158, 267, 272, 273, 364
 - multipersonales 68, 75
- egoísmo 223
- estratagemas 195, 196, 198, 202, 204, 205, 271, 273, 276, 280
- estrategia iii, 53-57, 60-62, 64, 66-69, 113, 122-128, 136, 138-143, 164, 170, 171, 179, 183, 216, 256-258, 264, 271-274, 276, 278, 280, 336-344, 363, 364
 - conjunta iii, 66, 67, 77, 79, 86, 90, 115, 122, 315, 337-339
 - cooperativa 62, 66-69, 74, 75, 82, 83, 85, 88, 91, 92, 95, 96, 101, 103, 106-108, 112-116, 119-121, 124-126, 128, 129, 135, 137, 140-143, 146-150, 155, 156, 158, 162, 163, 166, 217, 273, 274, 322, 325
 - dominante 67, 125, 126, 217, 222, 265, 268

mixta 62, 127
 no cooperativa 66-69, 75, 82, 85, 88, 91-96, 101, 103, 106-108,
 112-116, 119-121, 124-126, 128, 135, 140-142, 146-149, 150,
 155, 322
 pura 75

free-rider ver "aprovechado"

imparcialidad 90, 110, 179

interés propio iii, iv, 2, 136, 145-147, 157-171, 173, 176, 178-180,
 183, 186, 203, 208, 212, 218, 222-225, 227-233, 235-238,
 240, 242, 243, 254, 257, 259-262, 264, 267, 272, 274-276,
 279, 328, 340, 342-344, 351

irracionalidad 23, 24, 26-30, 186, 191-194, 199, 203, 262, 363

juegos

forma extensiva 51, 52
 forma normal 51, 52, 127, 128, 139
 intereses idénticos 57, 62, 122, 123, 130, 149, 159, 174, 217
 intereses mixtos 57, 63, 72, 87, 115, 123, 133, 149, 155, 158,
 159, 174, 217
 intereses opuestos 57-60, 62, 74, 217, 340
 reglas 37, 51, 52, 81, 122, 127,
 suma cero 59, 61, 65

maximización 14, 33, 37, 43, 45-49, 54-56, 62, 66, 81, 96, 109, 111,
 112, 145, 150, 161, 212, 218, 224, 227, 296-299, 302, 303,
 334

mejor respuesta 56, 57, 60, 61, 125, 126

moral i, ii, iii, iv, 79, 110, 139, 145-150, 158, 162-170, 175-183,
 227-234, 237-239, 242, 258-265, 270-275, 278-281, 288, 289,
 300, 336, 341-348, 352-359, 361, 362, 364, 365, 366-369,
 371

nivel de seguridad 60-62, 85, 86, 91

obediencia 113, 114, 116-118

 aquiescencia 116, 119

 plena 113, 116, 117, 241

 restringida 33, 48, 108, 113-118, 264, 330

óptimo 49, 63, 65, 66, 69, 72-74, 77, 87, 103, 113, 142, 143, 152,
 159, 217, 268, 269, 272, 275, 276, 313, 316, 319, 339

preferencias i, ii, iii, 10, 12-17, 19-32, 34, 35, 38-41, 43, 45, 48,
 50, 54, 57, 63, 64, 72-74, 133, 149, 156, 164, 173, 174-177,
 179-181, 186, 187, 189-199, 201-206, 209, 219, 220-222,
 224-234, 236-239, 247-255, 267-277, 294-299, 307, 310, 342
 morales 170, 175-177, 179, 227-234, 236-238, 254, 255, 261-263,
 267, 269, 270-272, 274, 275

- personales 129, 146, 159, 170, 171, 173-177, 180, 181, 218, 225-233, 236-239, 252, 254, 258, 261, 267, 269, 271, 272, 275, 344, 366
- primarias 25, 28-30, 196-199, 247
- secundarias 28, 197-199
- punto de encabalgadura 60, 61
- punto de inversión 42, 43
- racionalidad i, ii, iii, iv, 2, 13-17, 21, 23, 29, 33-35, 41-51, 54-56, 62, 67-70, 72, 78, 79, 87, 89, 105, 109, 117-119, 131, 133, 142, 144, 156, 161, 169, 171, 187, 189, 192, 199, 202-208, 210-212, 215, 218-222, 226, 234-238, 255, 256, 262, 265, 271-273, 275, 278, 281, 297-300, 302, 307, 329, 361-364
 - directa iii, 35, 36, 162, 180, 183-185, 236, 309
 - estratégica iii, 54-56, 61, 62, 74
 - indirecta iv, 27, 183-185, 202, 206, 211, 214, 250, 254, 264, 266, 269, 271, 272, 280, 300
 - paramétrica 54, 61
- salvaguarda lockeana 110, 145, 334
- satisfacción 27, 44, 45, 49, 299, 301, 303,
- status quo 75-78, 82, 85-87, 316, 321, 327, 330
- utilidad i, 11-15, 35, 57-67, 170, 174, 176, 195, 239, 252-254, 296, 301, 309-314,
 - cardinal 40, 43, 49, 58, 59, 297, 298, 307
 - esperada 41, 43, 54, 61, 62, 81, 82, 212, 307, 308, 365, 366
 - ordinal 297, 298